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

Chun Wang[1,2*] **Xiaojun Ye**[1*] **Xiaoran Pan**[1] **Zihao Pan**[3] **Haofan Wang**[4] **Yiren Song**[2,5†]
[1]Zhejiang University [2]Creatly.ai [3]Sun Yat-sen University [4]LibLib.ai [5]NUS
{chunwang0326,songyiren725}@gmail.com

Abstract

Recent advances in Visual Language Models (VLMs) have demonstrated exceptional performance in visual reasoning tasks. However, geo-localization presents unique challenges, requiring the extraction of multigranular visual cues from images and their integration with external world knowledge for systematic reasoning. Current approaches to geo-localization tasks often lack robust reasoning mechanisms and explainability, limiting their effectiveness. To address these limitations, we propose the **G**eo **R**eason **E**nhancement (**GRE**) Suite, a novel framework that augments VLMs with structured reasoning chains for accurate and interpretable location inference. The **GRE** Suite is systematically developed across three key dimensions: dataset, model, and benchmark. First, we introduce **GRE30K**, a high-quality geo-localization reasoning dataset designed to facilitate fine-grained visual and contextual analysis. Next, we present the **GRE** model, 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. Finally, we construct the **G**eo **R**eason **E**valuation Benchmark (**GREval-Bench**), a comprehensive evaluation framework that assesses VLMs across diverse urban, natural, and landmark scenes to measure both coarse-grained (e.g., country, continent) and fine-grained (e.g., city, street) localization performance. Experimental results demonstrate that **GRE** significantly outperforms existing methods across all granularities of geo-localization tasks, underscoring the efficacy of reasoning-augmented VLMs in complex geographic inference. Code and data will be released at `https://github.com/Thorin215/GRE`.

1 Introduction

Worldwide image geo-localization [40, 58] aims to predict the geographical coordinates of the shooting location based on any given photo taken anywhere on Earth. Unlike geo-localization within specific regions [25, 36, 52], global geo-localization, unrestricted to any specific region but covering the entire Earth, greatly unleashes the potential of geo-localization, which has significant applications across multiple domains, such as autonomous driving system positioning, social media image geo-tagging, and cultural heritage preservation. However, precise global-scale image geo-localization still faces substantial technical challenges due to the vast diversity of global geographical environments, visual ambiguity between similar locations, and the variability of shooting conditions including weather patterns, seasonal changes, and lighting conditions.

Geo-localization requires predicting the geographic coordinates of a photograph solely from the ground-view image. Extracting general geographical visual semantics is insufficient for the task, as

*The two authors contribute equally to this work.
†Corresponding author : songyiren725@gmail.com, Creatly.ai.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

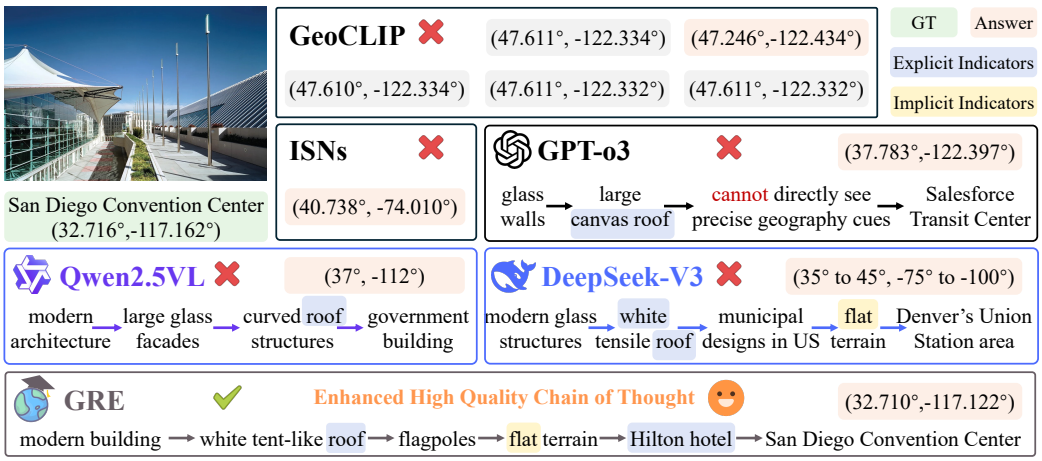


Figure 1: Performance comparison of our reasoning-based GRE versus traditional alignment-based approaches and MLLM baselines on image geo-localization.

two distant locations could potentially share similar image-level features. Instead, models need to **identify** and **reason** with geographically relevant visual elements from complex visual information. As illustrated in Fig. 1, when inferring the target location - San Diego Convention Center, the model is expected to jointly leverage explicit indicators such as the "white sail" roof design and implicit indicators such as flat terrain. However, existing approaches [21, 57] rely on data-driven cross-modal alignment strategies, which establish correspondences through large-scale annotated image-GPS pairs while neglecting the inherent logical relationships among fine-grained geographical indicators within images. In addition, models need to predict geographic coordinates for images captured at any location in the world. However, existing methods based on closed-domain assumptions either maintain a candidate database of GPS coordinates [21, 87] or images [32, 56, 61, 63, 90], or divide the entire geographical space into fixed grids for classification purposes [8, 19, 35, 40, 57], compromising the continuity of coordinate prediction. Thus, it is essential for image geo-localization models to possess the ability to predict **open-ended coordinates** without relying on candidate information, a feature that current methods inadequately address.

Recently, DeepSeek-R1 [12] has successfully applied Reinforcement Learning (RL) to induce the self-emergence of complex cognitive reasoning ability in LLMs. Image geolocalization is inherently a multi-step cognitive process that requires progressive reasoning - from identifying visual cues in images, to inferring geographical correlations among these cues, and ultimately determining specific locations. This progressive reasoning process aligns naturally with the sequential decision-making characteristics of RL. Through RL, models can learn to formulate optimal reasoning strategies based on identified visual features, gradually narrowing down potential geographical regions, and ultimately arriving at accurate location predictions, rather than simply relying on pre-established image-GPS correspondences. Unfortunately, this direct RL training is challenged, as it struggles to effectively guide MLLMs generating complex CoT reasoning in absence of large-scale, high-quality multimodal data and prolonged training [17]. What's more, fine-grained analysis of intermediate reasoning processes has proved beneficial for both evaluating and further improving models' reasoning capabilities [22, 65]. However, existing image geo-localization benchmarks [8, 13] focus solely on terminal prediction accuracy while ignoring reasoning quality assessment.

To address the aforementioned challenges, we propose **G**eo **R**eason **E**nhancement (**GRE**), a novel reasoning solution that integrates cold-start supervised fine-tuning and two-stage reinforcement learning training for worldwide image geolocalization. To facilitate the training process, we establish a geography reasoning dataset **GRE30k** by leveraging o3 to generate chain-of-thought demonstrations for geography seed questions. Our curated GRE30K consists of two sub-datasets: GRE30K-CoT, which contains format-standardized CoT content and answers refined through annotator filtering, and GRE30K-Judge, which comprises reasoning chain judgment tasks constructed through regular expression matching. GRE30k-CoT serves as a cold start dataset to establish basic reasoning capabilities of the base model. Then, we need to apply two-stage Group Relative Policy Optimization (GRPO) [12, 45] on a GRE30K-Judge and seed questions to enhance the model's reasoning capability.

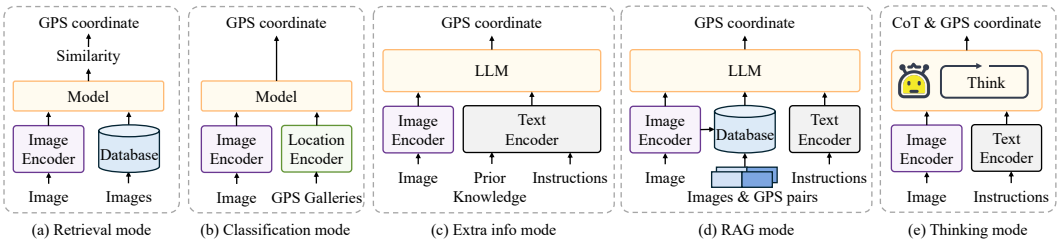


Figure 2: Summary of current image geo-localization model architectures.

Furthermore, to rigorously assess models' ability to leverage geographical visual cues for geo-localization and evaluate the quality of their reasoning chains, we develop a benchmark named **G**eo **R**eason **E**valuation Benchmark (**GREval-Bench**). Specifically, we design an automated pipeline to filter images containing geographical indicators and provide each image with a corpus of explicit and implicit geographical identifiers along with high-quality CoT annotations. We summarize the key contributions of our work as follows:

- We present **GRE**, a novel reasoning solution for the worldwide image geo-localization task. Our proposed methodology integrates cold-start initialization with a two-stage reinforcement learning training paradigm to effectively leverage geographical indicators within images and enable open-ended geolocalization.
- We introduce **GER30K**, comprising a high-quality CoT dataset and a judgement task dataset. We anticipate the dataset will benefit more future work for location-aware visual reasoning.
- Furthermore, to comprehensively evaluate the image geo-localization capability of the models, we develop **GREval-Bench**, consisting of higher quality images, CoT quality assessments, and a corpus of geographic indicators.

2 Related Work

Image Geo-localization. Image Geo-localization is an important task in computer vision [91–93], spatial data mining [82], and GeoAI [83]. As shown in Fig. 2, previous work in image geo-localization can be divided into four main modes: classification mode, retrieval mode, prior knowledge mode and RAG mode. (1) Retrieval mode treat the image geo-localization task as a retrieval problem, typically maintaining a database of images [32, 56, 61, 63, 89, 90] or a gallery of GPS coordinates [57]. They take the most similar images and GPS coordinates to the query image as the predicted values. However, maintaining a global-level image database or GPS gallery is infeasible. (2) Classification mode [8, 35, 40, 44, 58, 60] divide the entire earth into multiple grid cells and assign the center coordinates as predicted values. Models are then trained to classify the input image into the correct cell. However, if the actual location of the image is far from the center of the predicted cell, there can still be significant errors, even if the cell prediction is correct. (3) Prior Knowledge mode approaches [57] incorporate higher-level geographical information, such as continental-scale priors, to enhance performance. Nevertheless, this approach essentially provides partial solutions, contradicting the fundamental purpose of the task. (4) RAG mode [21, 87] leverage large language models by retrieving relevant image-GPS pairs as references to optimize predictions. While there are also some tries based on diffusion method like *Around the World* [11], with application of flow matching and diffusion [16, 28, 47, 78, 80, 81] . However, these approaches rely on establishing large-scale aligned databases. In contrast to existing global image geo-localization approaches, we propose a reasoning-based methodology that leverages both explicit and implicit geographical indicators within images to predict open-ended coordinate prediction. Recent advances in MLLMs have enabled novel approaches leveraging their reasoning capabilities for geographic inference. While some works [10, 27, 55]employ explicit reasoning chains, they lack systematic evaluation of reasoning quality. Complementary work has developed datasets [2, 49] and reinforcement learning frameworks [55] to enhance human-like geospatial reasoning.

Vision Language Models (VLMs). Models in the vein of GPT-4o [37] achieve excellent visual understanding ability by integrating both visual and textual data. This integration enhances the models' ability to understand complex multi-modal inputs and enables more advanced AI systems [26, 29,

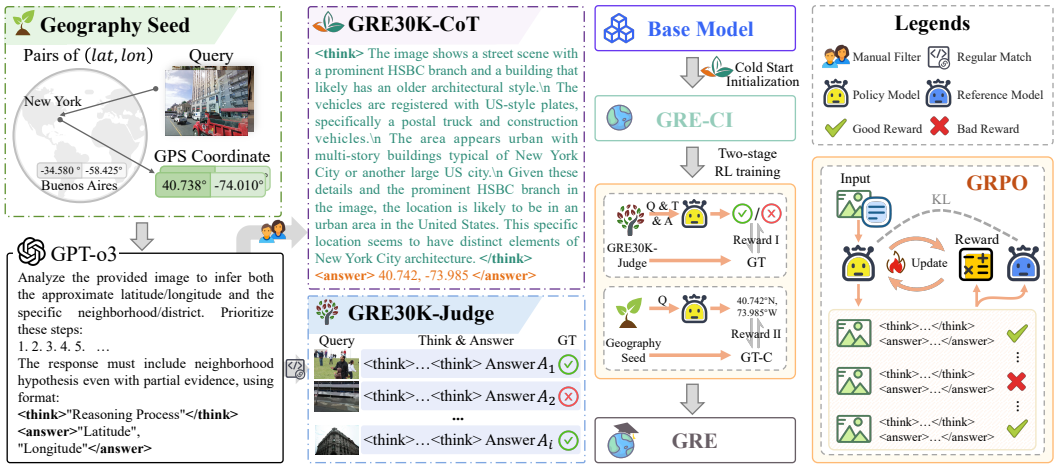


Figure 3: Overview of our GRE framework. The geographical reasoning pipeline begins with data preparation, incorporating automated CoT generation, regular expression matching, and manual filtering. Based on our constructed GRE30K dataset, we employ a post-training procedure that consists of supervised fine-tuning to learn reasoning patterns, followed by two-stage rule-based reinforcement learning to enhance image geo-localization reasoning capabilities.

59, 75] capable of processing and responding to both images and text. Generally, the training of LVLMs involves two steps: (a) pre-training and (b) post-training which contains supervised fine-tuning and reinforcement learning. Post-training is crucial in improving the model's response quality, instruction following, and reasoning abilities. While there has been significant research on using reinforcement learning to enhance LLMs during post-training [1, 5, 39, 42, 46, 50, 51, 66, 72, 85, 94], the progress for LVLMs has been slower. In this paper, we propose GRE-RL, which used GRPO-based reinforcement algorithms and verifiable reward during the post-training phase to enhance the model's visual perception and reasoning capabilities.

Reinforcement Learning. Recently, with the emergence of reasoning models like OpenAI's o1 [20] and Deepseek-R1 [12], the research focus in Large Language Models (LLMs) has increasingly shifted towards enhancing the models' reasoning capabilities through reinforcement learning (RL) techniques. Studies have explored improving LLMs' performance in reasoning tasks such as solving mathematical problems [4, 34, 45, 62, 70] and coding [18, 23, 74, 79]. A notable breakthrough in this area is Deepseek-R1-Zero [12], which introduced a new approach to achieving robust reasoning capabilities using RL merely, eliminating the supervised fine-tuning (SFT) stage. However, current research on RL-based reasoning has largely been confined to the language domain, with limited exploration of its application in multi-modal settings. For LVLMs, RL has primarily been used for tasks like mitigating hallucinations and aligning models with human preference [33, 48, 67–69, 71, 77, 84, 86]. Interpretable visual reasoning, once a longstanding challenge [14], now benefits from RL-finetuned LVLMs acting as decision agents [73]. Cutting-edge models like Kimi [53] demonstrate advanced capabilities, with research expanding beyond hallucination mitigation to core reasoning enhancement. However, there remains a significant gap in research focusing on enhancing reasoning and visual perception of Large Vision Language Models. To address this gap, our work uses a novel reinforcement fine-tuning strategy , applying verifiable rewards with GRPO-based [45] RL to visual geo-localization tasks. Our approach aims to improve the performance of LVLMs in processing various geo-localization tasks, especially when the high-quality fine-tuning data is limited.

3 Methodology

Fig. 3 illustrates the comprehensive reasoning pipeline of GRE. This method begins with a cold-start using a high-quality geo-localization Chain-of-Thought dataset, which initially teaches the base model to reason step-by-step following human-like patterns. Subsequently, we apply a two-stage reinforcement learning training to the cold-start initialized model GRE-CI to guide it towards adopting the correct geographical reasoning process, thereby enhancing the geo-localization reasoning capability in the final model GRE.

In the following sections, we first describe our approach to create a high-quality geo-localization reasoning dataset GRE30K in Section 3.1. Then we introduce our proposed Post-Training Strategy, comprising cold-start supervised fine-tuning (Section 3.2.1) and two-stage reinforcement learning training (Section 3.2.2). Correspondingly, our GRPO-based training strategy and two-stage reward function design will be described in Section 3.3.

3.1 GRE30K Construction

In this section, we present GRE30K, 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. Examples of the generated data are provided in Appendix A.1. While GWS15k [8] reveals Im2GPS3k [13]'s non-uniform distribution (with landmark repetition risks), our geographic filtering ensures clean evaluation.

Reasoning Process Generation. We make full use of the publicly available dataset MP16-Pro [21] with GPS coordinates. However, the source dataset only contains images, coordinates, and discrete geographical information including the corresponding county and state for each image, which are insufficient to train an MLLM. Our goal is to construct a CoT dataset that encompasses complex cognitive processes to facilitate our training strategy, enabling GRE to reason in a manner that closely resembles human cognitive patterns. Furthermore, GPT-o3 has demonstrated the capabilities in generating CoT reasoning that mirrors natural cognitive processes and has proven to have strong reasoning capability. Leveraging these insights, we employ GPT-o3 to generate image-CoT-coordinate triples through meticulously designed prompt templates. Please refer to Appendix A.2 for the detailed prompts for GPT-o3.

GRE30K-CoT. To address potential errors and mismatches in source CoT data, we combine automated filtering and manual verification to ensure the quality and reliability of the test data. Please refer to Appendix A.3 for more details. Finally, we collect 20k high-quality CoT samples. By acquiring CoT data in this manner, which closely mimics human cognitive behavior, reasoning processes exhibit natural and logical thinking.

GRE30K-Judge. In addition to standardizing the model's reasoning process through high-quality CoT data, we develop GRE30K-Judge, a judgment task dataset. This dataset is created by comparing extracted predictions with ground truth using threshold θ, labeling images as "Truth" or "False" accordingly. The resulting dataset is incorporated into reinforcement learning training, enabling the model to learn from both correct and incorrect reasoning patterns and thereby enhancing its geographical reasoning abilities. In total, we obtain 10k judgment samples.

3.2 Post-Training Strategy

To enhance visual reasoning capabilities, we introduce a three-stage post-training strategy consisting of cold-start initialization and two-stage rule-based reinforcement learning (RL). SFT stabilizes the model's reasoning process and standardizes its output format, while RL further improves generalization across various geo-localization tasks.

3.2.1 Cold-start Initialization

Leveraging the GRE30K-CoT dataset, we conduct SFT on a pretrained MLLM as the base MLLM for cold-start initialization. The MLLM after cold start initialization is named as GRE-CI. At this stage, the base MLLM had learned the complex reasoning mode from o3 [38]. Through SFT with the GRE30K-CoT dataset, the model standardize output format and establish a systematic reasoning framework. This critical phase facilitates the model's acquisition of high-quality structured reasoning patterns, thereby constructing a solid foundation for subsequent RL procedures.

3.2.2 Reinforcement Learning on the GRE-CI

Building upon the SFT-trained model, we employ rule-based reinforcement learning (RL) to optimize structured reasoning and ensure output validity. Specifically, we define two kinds of reward rules inspired by R1 and update the model using Group Relative Policy Optimization (GRPO). The RL stage further encourages the model to generate reliable outputs and enhances its generalization

capabilities in geographical reasoning tasks. Please refer to Appendix C.1 for more details about the two-stage RL training pipeline.

Rule-Based Rewards. We define two kinds of reward rules that evaluate the generated answers from two perspectives:

- **Accuracy Reward**: The accuracy reward rule evaluates the correctness of the final answer by extracting final answer via regular expressions and verifying them against the ground truth. For image geo-localization task, the final answer must be provided in a specified format to enable reliable rule-based verification. In **RL stage I**, given an input image along with its CoT and predicted answer, the model evaluates the correctness of both the reasoning process and the final answer. The model receives a reward score of $r_i = 1$ only if the generated final result aligns with the ground truth; otherwise, it receives a score of $r_i = 0$. In **RL stage II**, where the model directly predicts coordinates based on the input image, the reward is determined by the threshold metric θ.
- **Format Reward**: In order to ensure the existence of the reasoning process, the format reward rule requires that the response must follow a strict format where the model's reasoning is enclosed between `<think>` and `</think>`. A regular expression ensures the presence and correct ordering of these reasoning markers. What's more, `<answer>` and `</answer>` are used to ensure model have given a answer.

3.3 Group Relative Policy Optimization

We employ GRPO to achieve balanced integration of consistent policy updates and robust reward signals in a controlled manner. For each token in the generated output, GRPO first compute the log probabilities under both the new policy (π_θ) and a reference policy (π_{ref}). It then calculates the probability ratio and clips it to the range $[1-\epsilon, 1+\epsilon]$ to constrain policy updates and avoid divergence. The normalized reward (treated as an advantage estimate) is subsequently used in a PPO-style loss function, combining policy optimization with KL-divergence (weighted by β) regularization:

$$\mathcal{L}_{\text{clip}} = -\mathbb{E}\Big[\min\Big(\text{ratio}_t \cdot \text{Adv}_t,\ \text{clipped_ratio}_t \cdot \text{Adv}_t\Big)\Big]. \tag{1}$$

$$\begin{aligned}\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}\Big[\ &\min\Big(\text{ratio}_t \cdot \text{Adv}_t,\ \text{clipped_ratio}_t \cdot \text{Adv}_t\Big) \\ &- \beta \cdot \text{KL}\Big(\pi_\theta(y \mid x),\ \pi_{\text{ref}}(y \mid x)\Big)\Big].\end{aligned} \tag{2}$$

Here, Adv_t denotes the advantage function, capturing how much better (or worse) a particular action is compared to a baseline policy value.

Compared to other methods, the GRPO clipping mechanism prevents extreme policy shifts, while the KL regularization keeps the updated policy aligned with the baseline. This combination ensures that our model integrates rule-based rewards efficiently without compromising training stability. Subsequently, we will introduce the reward function R adopted for second-stage(Eq. (4)) and third-stage(Eq. (5)).

$$d = \text{geodesic}\big((\phi_{\text{pred}}, \lambda_{\text{pred}}), (\phi_{\text{true}}, \lambda_{\text{true}})\big) \tag{3}$$

$$R_{\text{yes/no}}(y_{\text{pred}}, y_{\text{true}}) = \begin{cases} 1.0 & \text{if } \mathcal{E}(y_{\text{pred}}) = \mathcal{E}(y_{\text{true}}) \\ 0.0 & \text{otherwise} \end{cases} \tag{4}$$

$$R_{\text{geo}}(y_{\text{pred}}, y_{\text{true}}) = \begin{cases} \frac{2}{1+\exp(d/\theta)} & \text{if } \mathcal{V}(y_{\text{pred}}, y_{\text{true}}) = \text{True} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Here, θ denotes the threshold, it is used as a factor to control the range of reward in this reward function Eq. (5). $\mathcal{E}$ mean the boolean value of the prediction and $\mathcal{V}$ mean the values of prediction and ground truth are valid.

4 GREval-Bench

To comprehensively evaluate the image geo-localization capability of the models, we develop a geographical reasoning benchmark named **GREval-Bench**. Existing benchmarks [13, 54] are directly constructed from geotagged Flickr images without appropriate filtering. Specifically, these benchmarks contain numerous images that lack geographical relevance cues, such as portraits and object-focused photographs. The inclusion of such geographically uninformative samples compromises the validity of evaluation results. Moreover, these benchmarks primarily focus on final predictions while neglecting the evaluation of the entire CoT process. The CoT process reflects multiple aspects of geographical reasoning capabilities and serves as a critical medium for understanding models' reasoning patterns and limitations.

To address these challenges, we propose an semi-automated pipeline for geo-localization image filtering and CoT annotation generation in our GREval-Bench. Fig. 4 and Table 1 provide data statistics, respectively. Please refer to Appendix B.1 for more details of the GREval-Bench construction and evaluation pipeline. GREval-Bench comprises 3K triplets, each containing: (1) geographical inference images filtered through our pipeline, (2) a corresponding corpus of geographical indicators categorized into explicit and implicit types, with detailed subcategories presented in Appendix B.2, and (3) reference GPS coordinates and annotated key Chain-of-Thought steps, where step categories and partitioning follow [22]. Through our construction pipeline, we have enhanced both the image quality and complexity of the benchmark by eliminating noisy images lacking geographical indicators while increasing the proportion of samples that require reasoning based on implicit indicators. This improvement facilitates a more accurate assessment of models' geo-localization capabilities.

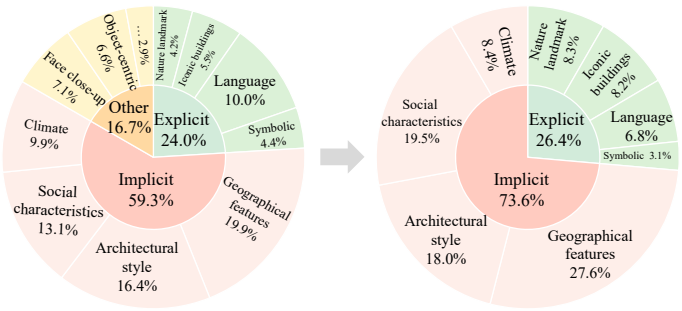


Figure 4: Indicators distribution of GREval-Bench.

Statistic	Number
Outdoor	2400
- natural scene	811
- artificial landscape	1138
- agricultural scene	58
- industrial scene	66
- road traffic	327
Indoor	600
- commercial premises	147
- offices	131
- transportation place	54
- cultural sites	148
- medical place	4
- entertainment venues	116

Table 1: Statistics of GREval-Bench.

As illustrated in Fig. 5, we instruct GPT-4o [37] to categorize each reasoning step into three categories: background information, image caption, and logical inference. We calculate the recall between background information and the corresponding geography corpus. Then, we employ RefCLIPScore [15] to evaluate the semantic alignment between image captions and visual content, and utilize BertScore [76] to assess the similarity between predicted and ground-truth logical inference steps. As these components are crucial for visual reasoning, we calculate CoT-quality by the follow equation (Eq. (6)).

$$\text{CoT-quality} = \frac{\text{Recall} + \text{RefCLIPS} + \text{BertS}}{3} \tag{6}$$

5 Experiment

Datasets and Evaluation details: We randomly sample 5% of MP-16 [24], a dataset containing 4.72 million geotagged images from Flickr [3], as geography seed datasets to construct our GRE30K. This dataset is strategically utilized across our three-stage training process: GRE30K-CoT, comprising 20k high-quality Chain-of-Thought examples curated by geography experts and standardized in format, serves for cold-start initialization; GRE30K-Judge, consisting of 10k CoT judgment tasks, is employed for Stage I reinforcement learning training and the remaining 170k seed datasets are utilized for Stage II reinforcement learning training. We test our trained model on Im2GPS3k [13] and Google World Streets 15k (GWS15k) [8]. To ensure a fair comparison with existing methods in the evaluation of Im2GPS3k, both our proposed model and transformer-based models are trained using only 5% of the MP-16 dataset. Follow the protocol followed in previous works [21, 57], we

[3] https://www.flickr.com/

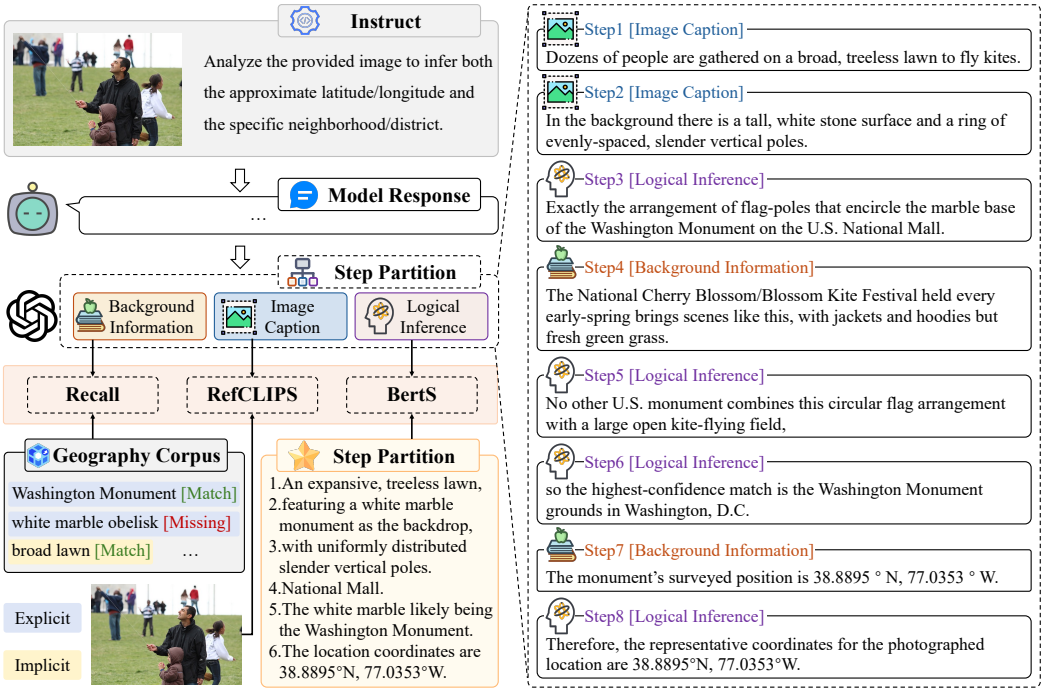


Figure 5: A detailed illustration of the evaluation pipeline.

report our results using a threshold metric. Given the predicted coordinates and the ground truths, this metric quantifies the percentage of predictions where the distance to the ground truth falls within specified thresholds (1km, 25km, 200km, 750km, and 2500km).

Implementation details: We adopt Qwen2.5-VL-7B as base model, the SFT experiments are conducted with a batch size of 128, a learning rate of 1e-5, and training over 1 epochs. Then, we perform RL on our dataset and experiment with training subsets of 10k for a single epoch each. All experiments are conducted with PyTorch and 8 NVIDIA H20(96G) GPUs.

5.1 Comparison with State-of-the-art methods

We perform a comparative analysis of GRE against worldwide Geo-Localization benchmarks, Im2GPS3k and GWS15k. The results on Im2GPS3k [13] and GWS15k [8] are shown in Table 2. In all metrics, our method surpasses the previous state-of-the-art (SOTA) 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. The results on additional geographical benchmarks are put in Appendix C.2, where we also observe a similar trend.

Moreover, our approach exhibits a large gain on the more challenging GWS15k dataset, surpassing the previous SOTA model with significant accuracy improvements of +0.2%, +1.0%, +2.0%, and +4.2% in the 1km, 25km, 200km and 2500km thresholds respectively. Our model achieves superior performance over previous state-of-the-art approaches while utilizing merely 5% of the data, compared to their use of the complete MP-16 dataset. The GWS15k contains samples that are uniformly sampled across the Earth and are not biased towards any specific geographic location. Moreover, the images in this dataset have a large distribution shift compared to the training set, making the geo-localization task tough and challenging for brute-force alignment approaches. Our substantial improvement can be attributed to effective reasoning that leverages both explicit and implicit geographical indicators within images.

Geo-localization in Vision-Language Models (VLMs) indeed highlights their ability to integrate world knowledge for inference—an emergent capability developed during training. To provide a comprehensive comparison, we have benchmarked both LLaVA-1.5 [31] and Molmo-D-7B [9] on the Im2GPS3k dataset, which use open-source training data.

Table 2: We compare the performance of GRE with the state-of-the-art methods on (a) Im2GPS3k [13] and (b) GWS15k [8] datasets. Our method yields consistent gains across datasets and different distance thresholds. † denotes transformer-based models. The asterisk (∗) signifies that for a direct comparison, GeoReasoner was prompted to output coordinates, which differs from its default city-name output format.

(a) Results on the Im2GPS3k [13] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
[L]kNN, $\sigma = 4$ [58]	7.2	19.4	26.9	38.9	55.9
PlaNet [60]	8.5	24.8	34.3	48.4	64.6
CPlaNet [44]	10.2	26.5	34.6	48.6	64.6
ISNs [35]	3.2	9.6	14.3	25.1	43.9
Translocator† [40]	7.6	20.3	27.1	40.7	63.3
GeoDecoder† [8]	5.7	10.3	21.4	28.9	38.6
GeoCLIP† [57]	10.8	31.1	48.7	67.6	83.2
GeoReasoner* [27]	0.2	1.6	2.1	3.9	6.8
GeoReasoner [27]	9.9	33.8	46.1	65.3	80.3
SeekWorld [55]	4.3	29.8	44.9	59.1	67.3
Qwen2.5-VL-7B [3]	3.2	16.6	28.0	42.1	53.0
LLaVA-v1.5-7B [30]	1.7	7.5	11.3	20.8	44.6
Molmo-D-7B [9]	2.1	9.8	19.6	36.3	55.7
Ours	**11.3**	**35.3**	**51.7**	**69.3**	**85.7**

(b) Results on the recent GWS15k [8] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
ISNs [35]	0.05	0.6	4.2	15.5	38.5
Translocator† [40]	0.5	1.1	8.0	25.5	48.3
GeoDecoder† [8]	0.7	1.5	8.7	26.9	50.5
GeoCLIP† [57]	0.6	3.1	16.9	45.7	74.1
GeoReasoner* [27]	0.01	0.01	2.3	10.9	18.0
GeoReasoner [27]	-	0.9	-	65.4	-
SeekWorld [55]	0.2	1.9	9.5	34.1	67.3
Ours	**0.9**	**4.1**	**18.9**	54.8	**78.3**

5.2 Performance on GREval-Bench

We compare our approach on GREval-Bench with the previous generalist models, including InternVL2.5 series [7], InternVL3 series [88], Qwen2.5-VL series [3]. 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. Table 3 presents the comparison results. Our approach achieves the leading average performance in various evaluation metrics while demonstrating more coherent reasoning processes that avoid local cognitive traps. Models with smaller parameter sizes like Qwen2.5VL-3B and InternVL3-2B exhibit significantly greater difficulty in extracting implicit cues compared to their larger counterparts. These models frequently commit errors in the early stages of CoT reasoning, compromising subsequent logical coherence. Fig. 6 illustrates a typical visual comparison.

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.

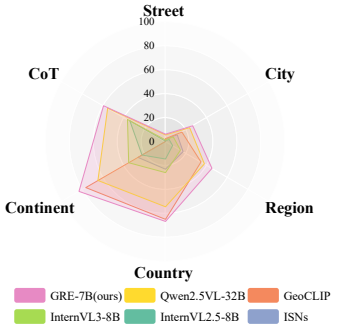


Figure 6: Visual demonstration of the performance of models.

5.3 Ablation Study

To evaluate the effectiveness of our training data and training strategies, we compare the model's performance under four distinct training strategies: (1) applying Cold-start Initialization on our dataset, (2) further optimizing the GRE-CI with RL stage I, (3) further optimizing the GRE-CI with RL stage II, and (4) further optimizing the GRE-CI with RL stage I and stage II. As illustrated in Table 4, the application of CI on our dataset significantly enhances the model's performance on both the coarse-grained (e.g., country, continent) and fine-grained (e.g., city, street) localization performance. For (2) and (3) , (3) reach a comparable performance and (2) dropped at some levels of granularity, attributed to the misalignment between training and test task (reward) types in Stage I. Overall, (4) demonstrates superior performance to (3) due to its more robust reasoning capabilities. We also conduct additional ablation study on larger scale model and other open source model in Appendix C.2, materials can be found in the repository.

Table 4: Ablation study on (a) Im2GPS3k [13] and (b) GWS15k [8] datasets.

(a) Results on the Im2GPS3k [13] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Qwen2.5-VL-7B	3.20	16.62	28.03	42.14	52.99
CI	7.77	29.30	44.78	62.43	78.81
CI + I	7.16	28.13	42.41	63.29	78.61
CI + II	10.96	36.11	52.17	67.26	83.32
CI + I + II	**11.33**	35.28	51.72	**69.33**	**85.67**

(b) Results on the recent GWS15k [8] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Qwen2.5-VL-7B	0.05	0.29	1.39	4.43	8.66
CI	0.45	2.17	12.91	37.58	61.83
CI + I	0.35	2.03	12.82	37.88	62.16
CI + II	0.88	3.91	18.69	55.61	78.03
CI + I + II	**0.91**	**4.13**	**18.86**	54.82	**78.28**

6 Conclusion

In this paper, we introduce a comprehensive framework for visual geo-localization reasoning, built upon a formalization approach that unifies data construction, model training, and evaluation. Our framework is designed to address the limitations of the current methods, enabling model to reason in geo-localization task. The ability of extracting of multigranular visual cues from images and integrating with external world knowledge will also inspire us in other domains of VLMs. This framework has led to the creation of the GRE dataset, a rich resource featuring detailed step-by-step reasoning annotations designed to enhance model training and evaluation on geo-localization task. The GRE model, trained using this framework, demonstrates strong geo-localization reasoning capabilities and exhibits robust generalization across a diverse range of scenes, from implicit scenes to explicit scenes. To further support the evaluation of geo-localization, we introduce GREval-Bench, a comprehensive benchmark that rigorously assesses model performance across various geospatial scenario. Our extensive experiments validate the effectiveness of our approach, showing significant improvements over state-of-the-art open-source models.

Acknowledgment

We would like to acknowledge the authors of R1-Onevision for their insightful responses to our technical questions. This work was fully supported by and is affiliated with Creatly.ai.

References

[1] Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. arXiv preprint arXiv:2311.18232, 2023.

[2] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, Constantin Aronssohn, Nacim Bouia, Stephanie Fu, Romain Loiseau, Van Nguyen Nguyen, Charles Raude, Elliot Vincent, Lintao Xu, Hongyu Zhou, and Loic Landrieu. Openstreetview-5m: The many roads to global visual geolocation. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 21967–21977, 2024. doi: 10.1109/CVPR52733.2024.02074.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[4] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. arXiv preprint arXiv:2403.17297, 2024.

[5] Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. In ICLR, 2023.

[6] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. arXiv preprint arXiv:2405.16473, 2024.

[7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271, 2024.

[8] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23182–23190, 2023.

[9] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 91–104, 2025. doi: 10.1109/CVPR52734.2025.00018.

[10] Zhiyang Dou, Zipeng Wang, Xumeng Han, Guorong Li, Zhipei Huang, and Zhenjun Han. Gaga: Towards interactive global geolocation assistant. arXiv preprint arXiv:2412.08907, 2025.

[11] Nicolas Dufour, David Picard, Vicky Kalogeiton, and Loic Landrieu. Around the world in 80 timesteps: A generative approach to global visual geolocation. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 23016–23026, 2025. doi: 10.1109/CVPR52734.2025.02143.

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.

[13] James Hays and Alexei A. Efros. Im2gps: estimating geographic information from a single image. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, 2008. doi: 10.1109/CVPR.2008.4587784.

[14] Feijuan He, Yaxian Wang, Xianglin Miao, and Xia Sun. Interpretable visual reasoning: A survey. Image and Vision Computing, 112:104194, 2021. ISSN 0262-8856.

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718, 2021.

[16] Dingbang Huang, Wenbo Li, Yifei Zhao, Xinyu Pan, Yanhong Zeng, and Bo Dai. Psdiffusion: Harmonized multi-layer image generation via layout and appearance alignment. arXiv preprint arXiv:2505.11468, 2025.

[17] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749, 2025.

[18] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2.5-coder technical report. arXiv preprint arXiv:2409.12186, 2024.

[19] Mike Izbicki, Evangelos E Papalexakis, and Vassilis J Tsotras. Exploiting the earth's spherical geometry to geolocate images. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II, pages 3–19. Springer, 2020.

[20] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. arXiv:2412.16720, 2024.

[21] Pengyue Jia, Yiding Liu, Xiaopeng Li, Xiangyu Zhao, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, and Dawei Yin. G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. Advances in Neural Information Processing Systems, 37:53198–53221, 2024.

[22] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. arXiv preprint arXiv:2502.09621, 2025.

[23] Fangkai Jiao, Geyang Guo, Xingxing Zhang, Nancy F Chen, Shafiq Joty, and Furu Wei. Preference optimization for reasoning with pseudo feedback. arXiv preprint arXiv:2411.16345, 2024.

[24] Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. The benchmarking initiative for multimedia evaluation: Mediaeval 2016. IEEE MultiMedia, 24(1):93–96, 2017.

[25] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5374–5384, 2022.

[26] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024.

[27] Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In International Conference on Machine Learning (ICML), 2024.

[28] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2023.

[29] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2024.

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, Advances in Neural Information Processing Systems, volume 36, pages 34892–34916. Curran Associates, Inc., 2023.

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In NeurIPS, 2023.

[32] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5624–5633, 2019.

[33] Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Mia-dpo: Multi-image augmented direct preference optimization for large vision-language models. arXiv preprint arXiv:2410.17637, 2024.

[34] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. arXiv preprint arXiv:2401.08967, 2024.

[35] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In Proceedings of the European conference on computer vision (ECCV), pages 563–579, 2018.

[36] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In Proceedings of the IEEE international conference on computer vision, pages 3456–3465, 2017.

[37] OpenAI. Hello gpt-4o, 2024. URL `https://openai.com/index/hello-gpt-4o/`.

[38] OpenAI. Openai o3 and o4-mini system card, 2025. URL `https://https://openai.com/index/introducing-o3-and-o4-mini/`.

[39] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In NeurIPS, 2022.

[40] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. In European Conference on Computer Vision, pages 196–215. Springer, 2022.

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PmLR, 2021.

[42] Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In ICLR, 2023.

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 815–823, 2015.

[44] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In Proceedings of the European Conference on Computer Vision (ECCV), pages 536–551, 2018.

[45] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, YK Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300, 2024.

[46] Charlie Victor Snell, Ilya Kostrikov, Yi Su, Sherry Yang, and Sergey Levine. Offline RL for natural language generation with implicit language q learning. In ICLR, 2023.

[47] Yiren Song, Danze Chen, and Mike Zheng Shou. Layertracer: Cognitive-aligned layered svg synthesis via diffusion transformer. arXiv preprint arXiv:2502.01105, 2025.

[48] Yiren Song, Cheng Liu, and Mike Zheng Shou. Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. arXiv preprint arXiv:2502.01572, 2025.

[49] Zirui Song, Jingpu Yang, Yuan Huang, Jonathan Tonglet, Zeyu Zhang, Tao Cheng, Meng Fang, Iryna Gurevych, and Xiuying Chen. Geolocation with real human gameplay data: A large-scale dataset and human-like reasoning framework. arXiv preprint arXiv:2502.13759, 2025.

[50] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In NeurIPS, 2022.

[51] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In ACL, 2024.

[52] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In proceedings of the IEEE/CVF international conference on computer vision, pages 12105–12115, 2021.

[53] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinhao Li, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yuhao Dong, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, Ziwei Chen, and Zongyu Lin. Kimi-vl technical report. arXiv preprint arXiv:2504.07491, 2025.

[54] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. Communications of the ACM, 59 (2):64–73, 2016.

[55] Kaibin Tian, Zijie Xin, and Jiazhen Liu. SeekWorld: Geolocation is a natural RL task for o3-like visual clue-tracking. `https://github.com/TheEighthDay/SeekWorld`, 2025. GitHub repository.

[56] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3608–3616, 2017.

[57] Vicente Vivanco Cepeda, 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, 36:8690–8701, 2023.

[58] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In Proceedings of the IEEE international conference on computer vision, pages 2621–2630, 2017.

[59] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.

[60] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, pages 37–55. Springer, 2016.

[61] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In Proceedings of the IEEE International Conference on Computer Vision, pages 3961–3969, 2015.

[62] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. arXiv preprint arXiv:2409.12122, 2024.

[63] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. Advances in Neural Information Processing Systems, 34:29009–29020, 2021.

[64] Lele Yang, Muxi Diao, Kongming Liang, and Zhanyu Ma. Grpo for llava. `https://github.com/PRIS-CV/GRPO-for-Llava`, 2025.

[65] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615, 2025.

[66] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In ICLR, 2023.

[67] Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. Mmad: Multi-modal movie audio description. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 11415–11428, 2024.

[68] Xiaojun Ye, Guanbao Liang, Chun Wang, Liangcheng Li, Pengfei Ke, Rui Wang, Bingxin Jia, Gang Huang, Qiao Sun, and Sheng Zhou. M4bench: A benchmark of multi-domain multi-granularity multi-image understanding for multi-modal large language models. In International Joint Conference on Artificial Intelligence, 2025.

[69] Xiaojun Ye, Chun Wang, Yiren Song, Sheng Zhou, Liangcheng Li, and Jiajun Bu. Focusedad: Character-centric movie audio description. arXiv preprint arXiv:2504.12157, 2025.

[70] Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. Internlm-math: Open math large language models toward verifiable reasoning. arXiv preprint arXiv:2402.06332, 2024.

[71] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. RlHF-V: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In CVPR, 2024.

[72] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. IJCV, 2024.

[73] Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In NeurIPS, 2024.

[74] Kechi Zhang, Ge Li, Yihong Dong, Jingjing Xu, Jun Zhang, Jing Su, Yongfei Liu, and Zhi Jin. Codedpo: Aligning code models with self generated and verified source code. arXiv preprint arXiv:2410.05605, 2024.

[75] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. arXiv preprint arXiv:2407.03320, 2024.

[76] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019.

[77] Yu Zhang, Changhao Pan, Wenxiang Guo, Ruiqi Li, Zhiyuan Zhu, Jialei Wang, Wenhao Xu, Jingyu Lu, Zhiqing Hong, Chuxin Wang, et al. Gtsinger: A global multi-technique singing corpus with realistic music scores for all singing tasks. Advances in Neural Information Processing Systems, 37:1117–1140, 2024.

[78] Yu Zhang, Wenxiang Guo, Changhao Pan, Zhiyuan Zhu, Tao Jin, and Zhou Zhao. Isdrama: Immersive spatial drama generation through multimodal prompting. arXiv preprint arXiv:2504.20630, 2025.

[79] Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. o1-coder: an o1 replication for coding. arXiv preprint arXiv:2412.00154, 2024.

[80] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8069–8078, 2024.

[81] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. arXiv preprint arXiv:2503.07027, 2025.

[82] Zijian Zhang, Xiangyu Zhao, Qidong Liu, Chunxu Zhang, Qian Ma, Wanyu Wang, Hongwei Zhao, Yiqi Wang, and Zitao Liu. Promptst: Prompt-enhanced spatio-temporal multi-attribute prediction. In CIKM.

[83] Sen Zhao, Wei Wei, Ding Zou, and Xianling Mao. Multi-view intent disentangle graph networks for bundle recommendation. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, volume 36, pages 4379–4387, 2022.

[84] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:2311.16839, 2023.

[85] Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn rl. In ICML, 2024.

[86] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. arXiv preprint arXiv:2402.11411, 2024.

[87] Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2749–2754, 2024.

[88] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479, 2025.

[89] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3640–3649, 2021.

[90] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1162–1171, 2022.

[91] Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Qidong Liu, Yongchao Ye, Wei Chen, Zijian Zhang, Xuetao Wei, and Yuxuan Liang. Controltraj: Controllable trajectory generation with topology-constrained diffusion model. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24, page 4676–4687.

[92] Yuanshao Zhu et al. Difftraj: Generating gps trajectory with diffusion probabilistic model. In Proceedings of the 37th Annual Conference on Neural Information Processing Systems, 2023.

[93] Yuanshao Zhu et al. Synmob: Creating high-fidelity synthetic gps trajectory dataset for urban mobility analysis. In Proceedings of the 37th Annual Conference on Neural Information Processing Systems, NeurIPS'23 (Dataset and Benchmarks Track), 2023.

[94] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. arXiv:1909.08593, 2019.

A More Details of GRE30K

A.1 Example Illustrations of GRE30K

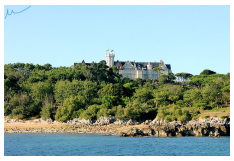
We provide several examples to illustrate the reasoning data in our GRE30K dataset, including the high-quality CoT data in Fig. 7 and the judgment data in Fig. 8.

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>**

GT: (40.746°, -73.990°)

<think> 1. Street-level inspection shows awning-style storefront signs stacked vertically – a common sight on the narrow, densely merchandised blocks of lower Manhattan where many import/wholesale businesses operate. // 2. Text clues: "Trading of NY, Inc.", "Handbag & Ladies Goods", "Empire Trading Company NY Inc." – the wording and emphasis on low-margin fashion accessories match the Canal Street / Chinatown wholesale district in New York City. // 3. No palm trees or desert flora; climate appears temperate with no visible vegetation, consistent with northeast U.S. city canyons. // 4. Terrain looks flat; Manhattan's southern half is largely level at sea level, matching the view. // 5. Vehicle is a graffiti-covered box truck typical of NYC curbside deliveries; heavy tagging is common in Lower Manhattan freight streets. **</think>**
<answer> "Latitude": "40.717°", "Longitude": "-73.999°" **</answer>**



GT: (43.464°, -3.799°)

<think> 1. Architecture: Grey-stone palace with steep slate roofs, dormers and a single crenellated tower flying two flags. This matches the early- 20th-century Palacio de la Magdalena used by the Spanish royal family in Santander. // 2. Street-level / shoreline context: Building sits on a wooded headland above a small sandy cove and jagged limestone rocks – characteristic of the Cantabrian coast of northern Spain. // 3. Textual clues: Though no close signage is readable, the dual flags on the tower are consistent with Spain's national flag plus the regional flag of Cantabria that normally fly over this palace. // 4. Vegetation & biome: Mix of maritime pine, Holm oak and lush Atlantic broad-leaf trees fits the temperate oceanic biome along Spain's Bay of Biscay rather than the drier Mediterranean coast. // 5. Terrain: Low cliff headland with gentle rise; this matches the topography of the La Magdalena Peninsula plotted on Santander nautical charts. // 6. Urban context/transport: Absence of tall urban waterfront buildings and presence of a heritage palace in parkland is consistent with the dedicated public park and university conference area of the La Magdalena neighbourhood rather than a dense downtown waterfront. **</think>**
<answer> "Latitude": "43.475°", "Longitude": "-3.767°" **</answer>**

Figure 7: Three examples to show CoT data in GRE30K-CoT.

GT Coordinate: (45.418°, -75.708°)

o3 Answer: (45.425°, -75.679°)

Judgment GT: True

Think:

1. 'Architecture & street furniture' – Gothic-Revival stone church with a *crossing* spire (set on the roof-ridge rather than at the front), stepped buttresses capped by finials, and a raised entrance reached by short exterior steps. That composition is typical of late-19th-century Anglican churches built of local limestone in Ontario/Quebec, rather than the brick or wood more common on the Prairies or in the U.S. Midwest. The unpaved road and plank sidewalk, together with the single utility pole carrying half-dozen open-wire telephone/telegraph lines, date the scene to c. 1890-1905 and suggest a medium-sized Canadian city that was just beginning to modernise. // 2. 'Textual clues' – No signage or licence plates are visible; only the negative number scratched in the emulsion (a hallmark of William J. Topley's Ottawa studio). // 3. 'Vegetation / biome' – Young deciduous street trees (likely maple/elm) and mown turf match the Great-Lakes–St-Lawrence mixed-forest zone (~43-46 ° N). // 4. 'Terrain' – Level ground with no hills on the skyline is consistent with the flat river-terrace neighbourhoods just east of Parliament Hill in Ottawa. // 5. 'Transport context' – Absence of street-rail tracks or heavy commercial façades suggests a residential district rather than the downtown retail core. Pulling these threads together, the building most closely resembles St Alban's Anglican Church (built 1867-74) in Ottawa's historic *Sandy Hill* district. That puts it just east of the Rideau Canal, a few blocks from the present-day University of Ottawa campus.

GT Coordinate: (40.724°, -73.901°)

o3 Answer: (33.773°, -118.121°)

Judgment GT: False

Think:

1. Street-level analysis: Mid-century modern A-frame church with mint-green steel ribs, full-glass façade, and a free-standing tapered brick campanile. That combination closely matches photographs of the Assumption of the Blessed Virgin Mary 'Greek Orthodox Church' built in 1959 in Long Beach, California. // 2. Textual clues: Over the doors, the inscription repeats the Greek word ΑΓΙΟΣ ("Holy"), confirming Greek Orthodox affiliation and pointing to an English-speaking city with a sizeable Greek community on the U.S. West Coast. // 3. Vegetation/biome: Trimmed box hedges, ornamental trees and warm-season turf are typical of a Mediterranean-type coastal climate (dry summers, mild winters) such as coastal Southern California (33-34 °N). // 4. Terrain: Site is flat with no visible foothills in the immediate backdrop—consistent with the coastal plain of Long Beach rather than the hillier Los Angeles basin edges. // 5. Urban context: Concrete sidewalks, U.S.-style streetlamps and adjacent low brick parish hall fit a post-war suburban neighborhood pattern common in East Long Beach (Park Estates/Los Altos district) where the documented church stands.

Figure 8: Two examples to show Judgment data in GRE30K-Judge. Red option indicates the wrong reasoning steps.

A.2 Detailed prompt for GPT-o3

Please refer to Fig. 9 for more details.

Instruction Prompt

Analyze the provided image to infer both the approximate latitude/longitude and the specific neighborhood/district. Prioritize these steps:
1. Identify architectural patterns, street furniture, and unique infrastructure
2. Decode textual clues (signage/license plates) for language or local references
3. Cross-reference vegetation types with regional biomes
4. Match terrain contours to topographic maps
5. Analyze transportation modes (vehicles/tracks) for urban context
The response must include neighborhood hypothesis even with partial evidence, using format:
<think>"Reasoning Process"**</think>**
<answer>"Latitude", "Longitude"**</answer>**
You must answer the question in this format, whether you have obtained the latitude and longitude through reasoning or not. If not, provide the closest possible values.

GT: (30.321°, 35.472°)

GPT-o3

<think>
1. Street-level cues: Rose-colored, carved sandstone façade with classical columns and lintels is iconic of the Nabataean monuments inside Petra (e.g., the Treasury/Al-Khazneh). No modern street furniture; setting is archaeological rather than urban.
2. Textual clues: None visible, but the absence of Latin alphabet signage and the presence of a traditionally dressed Bedouin guard/guide are consistent with tourist areas in southern Jordan.
3. Vegetation/biome: Arid, canyon-like walls of sandstone with no greenery point to a hot-desert biome of the Levant/Arabian plateau.
4. Terrain: Sheer sandstone walls in a narrow gorge match the Siq and surrounding carved monuments of Petra. Elevation ~800-900 m.
5. Transportation context: No vehicles—tourists in Petra traverse on foot or by animal, matching the pedestrian archaeological park setting.
</think>
<answer>
"Latitude": "30.322°", "Longitude": "35.452°"
</answer>

Figure 9: One example to illustrate the prompt for GPT-o3 to generate CoT data. The top block indicates the contexts including the image and instruction used to prompt o3, and the bottom block shows the response.

A.3 Review and Refinement Pipeline for GRE30K-CoT

Review and Refinement Pipeline for GRE30K-CoT. After the data generation process, we employ regular expression matching to filter out samples where the predicted coordinates deviate from the ground truth beyond a threshold θ. Notably, these filtered samples are not discarded but rather incorporated into GRE30K-Judge. To ensure the high quality of the generated samples, we apply manual verification after automated filtering. The process is conducted by three trained annotators with geographic-relevant professional backgrounds. The annotators examine and correct hallucinated image descriptions and inconsistent geographical reasoning in the CoT, ensuring that o3's output adheres to "`<think> </think><answer> </answer>`" format. Additionally, they maintain alignment between the reasoning process and the instruction structure.

Examples of Manual Filtering. As illustrated in Fig. 10, through a combination of regular expression matching and manual filtering, we enhance the quality of o3 generated Chain-of-Thought outputs, ultimately constructing a high-quality CoT dataset, GRE30K-CoT.

GT: (52.452°, -1.982°)

'The scene is an indoor community-hall with British-style stackable plastic chairs, noticeboards and pale interior finishes—common civic architecture in the UK Midlands. Crucially, the mug in the foreground bears the line "QUINTON ROAD WEST", a well-known arterial street in the Quinton suburb of Birmingham. That textual clue anchors the location. ~~Matching this to regional context (temperate broadleaf biome, gently rolling Midland terrain)~~ and the urban pattern of Quinton's local halls and churches leads to a neighbourhood hypothesis of Quinton, Birmingham, centred on Quinton Road West.',",52.46,-1.99

<think>
1. **Architectural context:** Generic Midlands civic infrastructure (stackable chairs/noticeboards/pale finishes) lacks regional specificity but aligns with UK community hall typology.
2. **Material culture clue:** Commercial mug displaying "QUINTON ROAD WEST" provides nominal locational anchoring to Birmingham's Quinton district, pending verification of souvenir distribution patterns.
3. **Urban morphology hypothesis:** Presumed correlation between community facilities' distribution (halls/churches) and Quinton's documented urban fabric remains conjectural without cadastral map alignment.
</think>
<answer>
"Latitude": "52.46°", "Longitude": "-1.99°"
</answer>

Figure 10: An illustrative example of Chain-of-Thought refinement and format normalization. The ~~red strikethrough text~~ denotes hallucinated content where the instructor model (o3) generated descriptions that are not actually present in the image.

B More Details of GREval-Bench

B.1 Detail of GREval-Bench Construction and Evaluation Pipeline

For image filtering, we construct a geographical reasoning corpus based on GRE30K-CoT, utilizing Named Entity Recognition (NER) to identify locations and architectural entities, and Semantic Role Labeling (SRL) to extract geographical reasoning patterns (e.g., "spire style $\rightarrow$ European church"). The geographical indicators in the corpus are then categorized into explicit and implicit types. Explicit indicators encompass artificial landmarks, natural geographical features, and textual symbols, while implicit indicators include architectural styles, urban planning patterns, social characteristics, and environmental characteristics. Please refer to Appendix B.2 for detailed sub-categories. We employ CLIP [41] to compute similarity scores between images and geography-relevant textual prompts from our geographical corpus (e.g., "base of Eiffel Tower", "Arabic text", "redwood forest"), retaining samples with high relevance scores. Subsequently, images with single facial regions occupying more than 50% of the area are removed through face detection [43]. The rule-filtered images then undergo manual verification, where annotators answer the question: "Can the approximate geographical location (country/city level) be inferred solely from this image?" Images are excluded if two or more out of three annotators respond negatively.

Inspired by previous CoT evaluation [6, 22, 65], we provide key steps annotation and reference GPS coordinate for all samples. We initially leverage o3 to generate the answer rationale. For the rationale, we provide both instructions and ground truth coordinates to o3. Subsequently, three geography domain annotators review and annotate key intermediate steps, utilizing o3's responses as reference. For cases where o3 fails to generate reasonable rationales, annotators develop geo-localization reasoning process independently.

B.2 Detailed Subcategories of Geographical Indicators

In the image geolocation task, geolocation indicators refer to the visual elements in the image that can directly or indirectly infer the geographic location. Table 5 shows the classification and specific examples of geolocation clues.

Table 5: Detailed subcategories of geographical indicators.

Type	Subcategory	Scenario
Explicit	nature landmark	• **Global/National Landmarks**: Eiffel Tower (Paris); Statue of Liberty (New York); Great Wall (Beijing) • **Regional Architecture**: Neuschwanstein Castle (Bavaria, Germany); Kiyomizu-dera Temple (Kyoto, Japan); Prague Astronomical Clock (Czech Republic) • **Unique Structures**: Bridges (Golden Gate Bridge); Ferris Wheel (London Eye); Religious Buildings (Mosque Domes, Gothic Church Spires)
Explicit	iconic buildings	• **Global/National Landmarks**: Eiffel Tower (Paris); Statue of Liberty (New York); Great Wall (Beijing) • **Regional Architecture**: Neuschwanstein Castle (Bavaria, Germany); Kiyomizu-dera Temple (Kyoto, Japan); Prague Astronomical Clock (Czech Republic) • **Unique Structures**: Bridges (Golden Gate Bridge); Ferris Wheel (London Eye); Religious Buildings (Mosque Domes, Gothic Church Spires)
Explicit	language	• **Language signs**: Language on road signs and store signs (Arabic $\rightarrow$ Middle Eastern; Cyrillic $\rightarrow$ Eastern European).
Explicit	symbolic	• **Administrative signs**: License plates (German license plates "D") • **Currency and flags**: Euro coins (European countries); Canadian maple leaf flag
Implicit	geographical features	• **Unique landforms**: Uyuni Salt Flats (Bolivia); Grand Canyon (USA); Guilin Karst landforms • **Vegetation types**: Cactus (desert areas); coconut trees (tropical coastal areas); birch trees (northern temperate zones) • **Water features**: Victoria Falls (Africa); Dead Sea (high salinity water bodies)
Implicit	architectural style	• **Architectural style**: Spanish colonial style (Mexico); neoclassicism (Washington, DC); earthen building (Fujian) • **Street characteristics**: Narrow cobblestone roads (European ancient towns); grid layout (Manhattan, New York); tricycles (Southeast Asian cities)
Implicit	social characteristics	• **Clothing and customs**: Kimono (Japan); Scottish plaid skirt; Indian sari • **Transportation**: Tunisian carriage; Venetian gondola; London red bus
Implicit	climate	• **Seasons and Weather**: Aurora (high latitudes); monsoon rainforest (rainy season in Southeast Asia); sandstorms (deserts in the Middle East)

C More Experiments

C.1 More Details on Training

Please refer to Fig. 11 and Fig. 12 for more details. During the training process, the threshold is continuously updated. If the model can stably maintain enough rewards at the current granularity level, the threshold is further refined to a finer granularity level.

C.2 Additional Main Results

We also conduct evaluations on the Google StreetView dataset(Table 6), where we observe similar performance trends. Additionally, we demonstrate the performance of our base model, Qwen-2.5VL series, on Im2GPS3k and GWS15k datasets(Table 7). The results align with our conclusions from the main results, further validating the effectiveness of our proposed training strategy. We also have compared our model on the OSV-5M [2] in Table 8, where our model emonstrates excellent performance. As *Around the World* demonstrates excellent performance, we conduct study on MP16 dataset, and materials can be found in the repository.

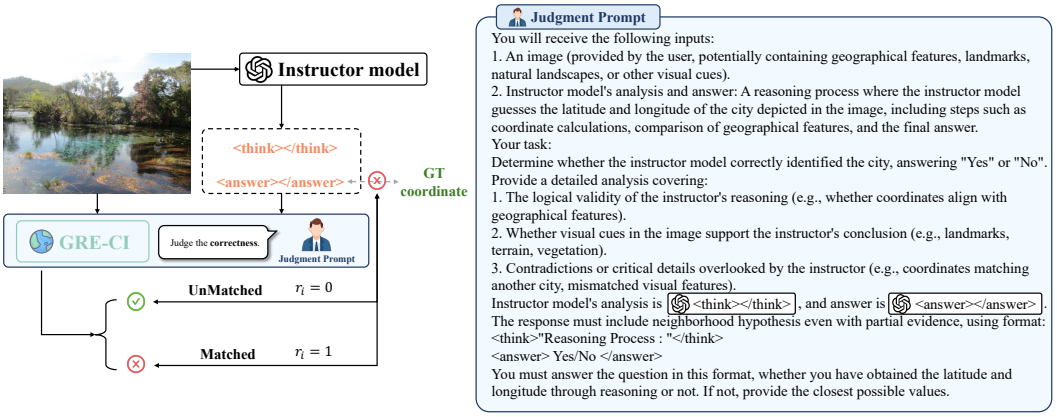


Figure 11: RL stage I training pipeline and Judgment Prompt.

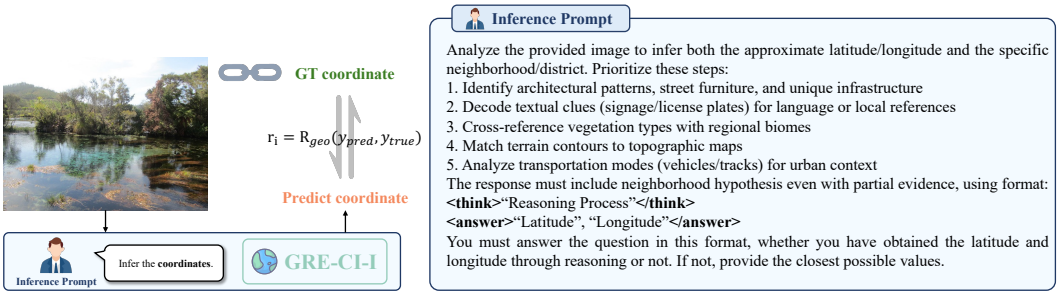


Figure 12: RL stage II training pipeline and Inference Prompt.

C.3 Additional Ablation Study

We also conduct additional ablation study on Qwen2.5VL-32B and LLaVA-v1.5-7B [64](Table 9), where we observe similar performance trends. The results demonstrate the efficacy and broad applicability of the proposed method.

C.4 Qualitative Results

In the supplementary materials, we provide additional visual examples illustrating the reasoning performance on the image geo-localization task. These examples demonstrate GRE's capability to generate remarkable chains of thought for accurate coordinate prediction in challenging scenarios.

D Limitations and Future Work

D.1 Limitations

The primary limitations of GRE include (1) substantial computational resource requirements, specifically utilizing 8 NVIDIA H20 GPUs for model training, and (2) the associated API costs for dataset generation. GeoCLIP requires 155.63 GFLOPs per inference. In comparison, our model requires 262.27 GFLOPs for the visual encoder and 24,117.47 GFLOPs for the language model, which corresponds to 13.0506 GFLOPs per token. All FLOPs are measured using the THOP package.

D.2 Future Work

Leveraging geo-localization reasoning capabilities, we can implement geographic information privacy identification and protection mechanisms. Furthermore, this approach can be extended through agent-based architectures that integrate reasoning capacities with tool invocation functionalities.

Table 6: Results on the Google StreetView dataset.

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Qwen2.5VL-3B	4.47	46.92	68.22	78.26	83.89
Qwen2.5VL-7B	7.99	61.00	70.42	83.20	85.56
Qwen2.5VL-32B	14.62	67.50	69.04	88.42	92.59
CI	15.53	64.25	74.46	<u>94.20</u>	<u>96.14</u>
CI + I	13.59	63.75	75.19	92.30	96.02
Ours	**18.15**	**71.01**	**75.36**	91.30	92.75

Table 7: We test the Qwen2.5VL series on (a) Im2GPS3k [13] and (b) GWS15k [8] datasets for reference here.

(a) Results on the Im2GPS3k [13] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Qwen2.5VL-3B	0.33	1.20	3.57	5.37	7.31
Qwen2.5VL-7B	3.20	16.62	28.03	42.14	52.99
Qwen2.5VL-32B	**6.47**	**25.12**	**40.96**	**59.87**	**75.32**

(b) Results on the recent GWS15k [8] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Qwen2.5VL-3B	0.02	0.17	0.41	2.14	6.70
Qwen2.5VL-7B	0.05	0.29	1.39	4.43	8.66
Qwen2.5VL-32B	**0.06**	**0.36**	**7.53**	**28.46**	**52.39**

D.3 Broader Impacts

The reasoning capacity improvement in geo-localization facilitates the extraction of multi-granularity geographic indicators from imagery, offering dual benefits for geospatial data mining applications and location privacy preservation frameworks.

E More Qualitative Results

We present additional visual examples to highlight the geographic reasoning performance. Fig. 13 displays more visual cases involving diverse locations. GRE is capable to generate explainable predictions with robust capabilities in these challenging scenarios. Furthermore, Fig. 14 and Fig. 15 provides comparisons with previous alignment-based methods and existing MLLMs with reasoning capabilities. Our approach exhibits superior image geo-localization results with implicit geographic indicators.

Table 8: Results on the recent OSV-5M [2] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km	dist Average Distance
Qwen2.5VL-7B	1.0	1.9	4.8	19.0	43.1	4942
Molmo-D-7B	0.7	1.1	1.3	7.2	32.1	6172
LLaVA-V1.5-7B	0.1	0.2	0.7	5.0	21.9	6895
SeekWorld	1.0	1.3	7.0	27.6	51.3	4326
SC Retrieval	-	<u>19.9</u>	<u>45.8</u>	73.4	-	1386
RFM $\mathcal{S}_2$	-	5.4	44.2	<u>76.2</u>	-	<u>1069</u>
Ours	**5.7**	9.7	35.57	72.53	**91.11**	1192

Table 9: More ablation study on (a) Im2GPS3k [13] and (b) GWS15k [8] datasets.

(a) Results on the Im2GPS3k [13] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
LLaVA-v1.5-7B	1.7	7.5	11.3	20.8	44.6
CI(LLaVA)	4.2	10.2	24.9	42.9	58.9
CI + II(LLaVA)	6.1	14.6	31.3	47.6	63.1
Qwen2.5-VL-32B	6.5	25.1	41.0	59.9	75.3
CI(Qwen)	8.1	31.4	46.5	69.7	81.1
CI + I(Qwen)	7.6	30.1	40.2	71.2	82.2
CI + I + II(Qwen)	**12.3**	**36.6**	**59.3**	**78.3**	**88.6**

(b) Results on the recent GWS15k [8] dataset

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Qwen2.5-VL-32B	0.06	0.36	7.5	28.5	52.4
CI	0.51	2.8	15.1	43.4	68.1
CI + I	0.42	2.3	13.9	43.6	68.1
CI + I + II	**0.97**	**4.9**	**20.1**	**57.0**	**81.3**

<think>
1. **Architectural patterns:** The image shows a large sculpture resembling a giant pencil surrounded by modern architecture, including a glass-roofed structure that could be part of a train station or transit hub.
2. **Textual clues:** The presence of Italian signage ("Yaris Il Piccolo Genio" and "APITALIA") suggests the location is in Italy. The design of the building and the style of the sculpture are consistent with Milan's urban landscape.
</think>
<answer>
"Latitude": "45.465°", "Longitude": "9.191°"
</answer>

<think>
1. **Textual clues:** Based on the signage of DiBruno Bros., which is known as 'The House of Cheese' and has been established since 1939, it is located in Philadelphia, Pennsylvania.
</think>
<answer>
"Latitude": ”39.952°", "Longitude": ”-75.165°"
</answer>

<think>
1. **Architectural patterns:** The image shows a street view of a city with tall buildings, including one that is notably the Willis Tower (formerly known as the Sears Tower), which is located in Chicago, Illinois.
</think>
<answer>
"Latitude": ”41.882°", "Longitude": ”-87.630°"
</answer>

Figure 13: Visual examples of GRE.

(10.784°, 106.681°)

GeoCLIP (17.966°, 102.608°)

(17.966°, 102.611°) (17.969°, 102.611°) (17.965°, 102.606°) (17.970°, 102.610°)

GPT-o3 (10.770°, 106.770°)

Architectural patterns – The covered five-foot-way with French-colonial arches, wrought-iron folding shop-gates, and the distinctive yellow interlocking-arc sidewalk tiles are characteristic of older shophouse blocks in central Ho Chi Minh City. Textual clues – The large café umbrella reads “TRUNG NGUYÊN Coffee”, a well-known Vietnamese brand; the small printed signs nearby use the Latin alphabet with diacritics typical of Vietnamese. No non-Vietnamese scripts are visible. 3. Vegetation/biome – Sparse ornamental plants and the bright sun point to a tropical low-latitude city; nothing indicates high elevation. Terrain contours – Entirely flat ground, consistent with the Saigon River delta plain. Transportation modes – Dense rows of scooters/motorbikes dominate the curb, matching commuting habits in Ho Chi Minh City’s core.

InternVL2.5 (21.03°, 105.84°)

Architectural Patterns and Infrastructure: The image shows a narrow street with shops on either side, typical of Southeast Asian urban areas. The buildings have a colonial architectural style, with arched doorways and tiled roofs. Textual Clues: The signage includes Vietnamese text, such as "Trung Nguyen," which is a well-known Vietnamese coffee brand. This suggests the location is in Vietnam. Vegetation and Biomes: The vegetation is not clearly visible, but the general environment suggests a tropical or subtropical climate, consistent with Vietnam. Terrain Contours: The street is flat, which is common in urban areas of Vietnam. Transportation Modes: Motorbikes are parked along the street, a common mode of transportation in Vietnam.

Qwen2.5VL (0°,-107°)

Architectural Patterns: The architecture suggests a style commonly found in markets across parts of Vietnam, such as Hanoi's Old Quarter. Textual Clues: There is signage visible on one umbrella which reads "TRUNG NGUYEN," indicating Vietnamese branding often associated with coffee shops and cafes throughout the country. Street Furniture & Unique Infrastructure: Red plastic stools typical of local markets can also hint at locations where similar items might still dominate public spaces today; Vegetation Types : No distinct plant species stand out here but rather seem consistent with tropical climates common among many Asian nations including Vietnam Terrain Contours : Not discernible enough given only ground level view Transportation Modes: Motorbikes parked along side suggest proximity to areas heavily reliant upon them - again indicative towards places popularizing motorbike culture alongside traditional retail setups.

GRE (10.823°, 106.630°)

The image depicts a market scene with a distinct Southeast Asian architectural style, likely in Vietnam given the signage on the umbrella which reads "Trung Nguyen." This is a well-known coffee brand in Vietnam. The market appears to be located in an urban area, possibly in Ho Chi Minh City or another major city in Vietnam.

Figure 14: Qualitative comparisons with previous alignment-based methods and existing MLLMs with reasoning capabilities. (Lat, Lon) denotes the ground truth coordinates, (Lat, Lon) denotes the models’ predicted answer, Indicator denotes the explicit indicator and Indicator denotes the implicit indicator. Notably, GeoCLIP generate five candidates coordinates and select the candidate with the maximum probability score as the answer.

(19.891°, -99.404°)

GeoCLIP

(-12.123°,-77.029°)

(-12.123°,-77.033°) (-12.045°,-77.031°) (-12.118°,-77.024°) (-12.122°,-77.027°)

GPT-o3 (19.433°,-99.133°)

Architectural patterns & street furniture – The heavy ash-grey volcanic-stone façade behind the kiosk, the wide concrete paver sidewalks, a Telmex pay-phone, and a glass “estanquillo” news-kiosk are hallmarks of Mexico’s capital-city historic core. Textual clues – The green fascia reads “MILENIO”, a national Mexican newspaper; the pay-phone logo is Telmex; the green-and-white Volkswagen Vocho is the livery used by Mexico City taxis until 2007. Terrain contours – Flat valley floor with no discernible upslope, matching the valley plain of the Distrito Federal. Transportation context – A vertical green totem traffic sign with a right-turn arrow is typical of the “Eje Vial” one-way grid in the Centro Histórico, and the traffic density fits the Cuauhtémoc borough hub.

Qwen2.5VL (-70° to -68 °, 19° to 23°)

Step 1: Architectural Patterns - The image shows modern buildings typical of many Latin American countries. Step 2: Textual Clues/Language References - There is signage that reads "Milenio," which suggests Spanish influence but does not specify an exact location within Spain as it's also used elsewhere like Mexico City where Milenio refers to a newspaper chain rather than being tied directly to any particular district name on its own without further context from license plate or other identifiers visible here." Step 3 & Step 4: Vegetation Types and Terrain Contours don't offer much additional information beyond suggesting temperate climate zones common across several parts globally including some areas in North America too! Step 5: Transportation Modes show cars parked along streets indicating urban setting; however again no clear indication narrows down precise city due solely upon vehicles present nor their design styles."

GRE (19.433°, -99.133°)

The image shows a street scene with cars parked along the curb, a newspaper stand, and a building with signage that reads "Milenio". This suggests a location where Spanish is spoken, as "Milenio" is a well-known Spanish newspaper. The architecture and street layout resemble those found in Latin America, particularly in Mexico City.

Figure 15: Qualitative comparisons.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize the contributions and scope in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have mentioned the limitations of our work in Appendix D.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We have contained the formula derivation process and the experimental results to prove the assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have introduced all the details of GRE in our paper. In addition, our submission also contains the materials to reproduce the main experimental results, including code, data, experiments settings, etc.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include the link to the GitHub repository in our paper to provide open access to the data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Section 5 and Appendix C.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We conducted repeated experiments and calculated the average to mitigate the variability of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have included the cost information in Appendix D.1

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We mention the societal impact on the Appendix D.3.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work does not have such risk.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper are well documented and is the documentation provided alongside the assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No potential risks are found in this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We use LLM to generating our dataset and we have described it in Section 3 and Appendix A.3.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.