Towards Interactive Global Geolocation Assistant

Zhiyang Dou[1,⋆], Zipeng Wang[1,⋆], Xumeng Han[2,⋆]
Guorong Li[3], Zhenjun Han[2,†], and Zhipei Huang[2]

[1] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences (UCAS), Beijing, China
[2] School of Electronic, Electrical and Communication Engineering, UCAS, Beijing, China
[3] School of Computer Science and Technology, UCAS, Beijing, China
hanzhj@ucas.ac.cn
⋆ Equal contribution. † Corresponding author.

Abstract. Global geolocation, the task of predicting precise coordinates from street-view imagery, is inherently plagued by visual ambiguity. Resolving such ambiguity necessitates a transition from static, one-shot predictions to an **interactive geolocation paradigm** driven by multi-turn deductive reasoning. However, most of existing geolocation models and general-purpose MLLMs fail to support this dynamic process, mainly due to the lack of interaction capabilities and the geographic knowledge gap. Driven by the imperative to actualize this interactive paradigm, we introduce **MG-Geo**, the first large-scale multimodal geolocation dataset explicitly structured for spatial reasoning. Comprising 4.87M geo-tagged Meta entries, 70K image-grounded Clue samples, and 73K multi-turn Dialog samples across 210 countries and territories, MG-Geo separates large-scale geographic alignment from reasoning-oriented supervision. Experiments demonstrate that GaGA achieves SOTA performance across several benchmarks. Notably, on the GWS15k dataset, it surpasses the strong Hybrid Model by 4.57% and 2.92% at the country and city levels, respectively, while securing the highest city-level accuracy (7.46%) on OSV-5M-test. More importantly, we formalize the "Similarity Trap"—a phenomenon where distributive visual features mislead static models—and demonstrate that GaGA effectively navigates this challenge through a *Tiered Interaction Protocol*. By dynamically integrating user-provided geographic anchors, GaGA achieves significant localization improvements. Our dataset is accessible via: https://huggingface.co/datasets/kendouvg/MG-Geo.

Keywords: Interactive geolocation · multimodal large language model · chain of thought

1 Introduction

Global geolocation aims to predict the exact location of any street-view image, with wide applications in security surveillance, emergency response, disease outbreak prediction, environmental monitoring, and tourism navigation [28, 30, 41].

This task requires integrating visual cues, such as road signs, architectural styles, climate, and vegetation, with geographic knowledge to accurately predict GPS coordinates or location labels. For images with landmarks or distinctive architecture, the location can be inferred by combining visual features with contextual knowledge. However, geolocation becomes more challenging in homogenous environments, such as highways or natural landscapes, where subtle geographic clues like road markings, license plate types, and signage must be relied upon.

Traditional street-view localization methods are generally categorized into retrieval-based and classification-based approaches. The retrieval-based methods [34, 39, 40, 43] match input images with similar ones from a geotagged database but are constrained by the diversity and completeness of the database. The classification-based methods [25,29,38] classify images into predefined regions based on visual features, but they lack interpretability and fail to provide explicit visual cues. In practical scenarios, geolocation is rarely a one-time, static process; it involves integrating and refining multiple sources of information iteratively through continuous interaction. Traditional geolocation models inherently lack interpretability and flexibility.

Multimodal Large Language Models (MLLMs) [1, 7, 23] offer a promising alternative due to their ability to synthesize cross-modal information and perform complex interpretive reasoning. This reasoning capacity is the key to unlocking an **interactive geolocation paradigm**, where spatial ambiguities can be dynamically resolved through multi-turn dialogue, hypothesis generation, and evidence accumulation. Yet, existing general-purpose MLLMs struggle to execute this interactive process, primarily due to a profound geographic knowledge gap [28]. They fail to establish robust, verifiable associations between granular geographic elements (e.g., specific road markings or vegetation patterns) and precise spatial coordinates.

A major bottleneck preventing this paradigm shift is the fundamental misalignment of existing training data. While conventional geolocation datasets, such as OSV-5M [2], are engineered exclusively for static, one-shot mapping, general multimodal datasets [5, 14] lack the necessary geographic granularity and administrative hierarchies. Driven by the imperative to actualize this interactive geolocation paradigm, we introduce **MG-Geo**, the first large-scale multimodal dataset explicitly structured to support multi-turn dialogue and deductive spatial reasoning. Featuring an *Interactive Reasoning Chain-of-Thought* strategy, MG-Geo contains 4.87M geo-tagged Meta entries, 70K image-grounded Clue samples, and 73K multi-turn Dialog samples, with the Clue and Dialog parts organized around eight categories of Geographic Element Cues. As depicted in Figure 1, it covers 210 countries and categorizes cues into eight critical dimensions—including road markings, vegetation, and linguistic signatures. By transitioning from isolated static labels to interconnected reasoning chains, MG-Geo provides the essential foundation for training MLLMs to perform dynamic, evidence-based geographic assistance.

Utilizing this resource, we propose the **G**lob**a**l **G**eolocation **A**ssistant (GaGA), an MLLM architected for high-precision, interactive spatial reasoning. GaGA

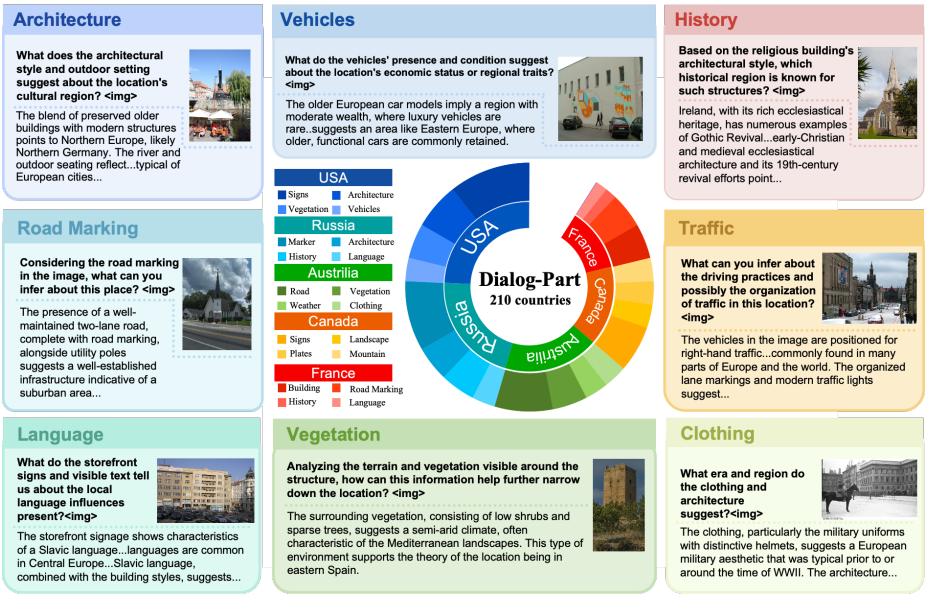

Fig. 1: Illustration of MG-Geo. Featuring diverse geographic scenes and visual cues, these images demonstrate the utility for training MLLMs to connect visual content with geographic locations and enrich their understanding of global environments.

excels not only in static benchmarks but also in navigating the **"Similarity Trap"**—a fundamental challenge where geographically widespread visual motifs (e.g., neoclassical architecture) lead to over-confident misclassifications. Our contributions are summarized as follows:

(1) We present **MG-Geo**, a large-scale multimodal geolocation dataset with 4.87M geo-tagged Meta entries, 70K image-grounded Clue samples, and 73K multi-turn Dialog samples, enabling MLLMs to move beyond static image tagging toward geographic reasoning and interaction.

(2) We develop **GaGA**, which establishes a new state-of-the-art (SOTA) across multiple benchmarks. On the GWS15k dataset, GaGA surpasses the strong Hybrid baseline by 4.57% and 2.92% at the country and city levels, respectively.

(3) We formalize and investigate the **Interactive Geolocation Paradigm**. Through a **Tiered Interaction Protocol**, we demonstrate how GaGA can effectively leverage user-provided anchors to resolve visual ambiguity.

2 Related work

2.1 Geolocation Datasets

In the domain of geolocation, the localizability of images within datasets is of paramount importance. Although the existing datasets, such as Im2GPS3k [35],

Table 1: Comparison between MG-Geo and existing datasets. MG-Geo is the first large-scale multimodal dataset featuring **expert-guided CoT**, **8-category geographic clues**, and **multi-turn interactive dialogues**.

Dataset	Size	Open	Source	Scope	Chain of Thought		Geo-Clues	Multi-turn Dialog
					AI-gen	Expert-guided		
MP-16 [32]	4.7M	✓	Web	Biased	✗	✗	✗	✗
OSV-5M [2]	5M	✓	Street-view	Global	✗	✗	✗	✗
Google Landmark V2 [37]	5M	✓	Landmark	Global	✗	✗	✗	✗
MP16-Reason [22]	33k	✓	Web	Biased	✓	✗	✓	✗
GRE30k [36]	30k	✓	Web	Biased	✓	✗	✗	✗
MG-Geo (ours)	**5M**	✓	**Street-view + Landmark**	**Global**	✓	✓	✓	✓

YFCC4K [33] and MP-16 [32] contain a large number of geotagged images, many of them are difficult to localize and exhibit distribution biases. GWS15k [8] mitigates distribution differences, and ensures that the images are authentic, localizable street views. OSV-5M [2] is the largest open-source collection of planet-scale, localizable street view images. The Google Landmark V2 [37] dataset contains globally distributed human-made and natural landmarks and showcase iconic landscapes.

We introduce MG-Geo, the first large-scale multimodal geolocation dataset comprising 4.87M geo-tagged Meta entries, 70K image-grounded Clue samples, and 73K multi-turn Dialog samples structured with an expert-guided Chain-of-Thought strategy to enhance the geospatial reasoning and interactivity of MLLMs. MG-Geo improves geographic coverage over existing reasoning datasets by combining OSV-5M street-view metadata with GLDv2 landmark imagery, while we visualize its global coordinate distribution and long-tail nature in the Appendix. It also incorporates well-structured global language knowledge with eight categories of detailed geographic element cues, providing a dataset that better reflects the complexity and diversity of real-world geolocation challenges.

2.2 Traditional Geolocation Methods

Traditional mainstream geolocation methods can be broadly categorized into two approaches: image-based retrieval and classification-based methods. Image-to-image retrieval techniques rely on dense image retrieval libraries, which perform well for localization tasks within small areas. However, the cost of constructing such retrieval libraries on a global scale is prohibitively high. When geolocation is treated as a classification task, categories can be defined based on administrative regions, divided into geocells according to specific rules, or discretized into latitude and longitude coordinates. TransLocator [25] employs images and semantic segmentation maps as inputs, facilitating interaction between two parallel branches after each Transformer layer and enabling multitask geolocation and scene recognition. GeoCLIP [34] introduces a location encoder and applies random Fourier feature representations to latitude and longitude coordinates. It utilizes the pretrained CLIP [27] visual encoder to represent images and aligns

them with the corresponding location features for localization. PIGEON [16] classifies within a self-created geocell and retrieves locations within clusters.

2.3 Generation-Based Geolocation Methods

Powerful MLLMs have injected new momentum into geolocation research. We refer to this emerging paradigm as generation-based geolocation, which has given rise to two predominant research directions. One dominant approach combines MLLMs with Retrieval-Augmented Generation (RAG) to enhance precision. This line of work includes Img2Loc [42], which reframes geolocation as a training-free text generation problem using CLIP-based retrieval; G3 [18] proposes a three-stage framework to mitigate challenges of visual semantics and data imbalance; and GeoRanker [19] introduces a distance-aware ranking framework to model spatial relationships among retrieved candidates. Another research direction focuses on enhancing the intrinsic reasoning abilities of MLLMs. For example, GeoReasoner [21] fine-tunes an MLLM for street view localization using a curated dataset of human reasoning patterns derived from geolocation games. Similarly, GLOBE [22] and GRE [36] develop a reasoning-oriented dataset from diverse social media images and use group-relative policy optimization to improve the generation of visual-clue-based rationales. Distinguished from prior work, GaGA is trained on a planet-scale dataset and introduces a novel interactive reasoning paradigm via dialogue.

3 MG-Geo Dataset

In this paper, we introduce MG-Geo, a novel dataset encompassing a diverse array of geographic element cues, including architecture, environment, landmarks, and climate across various countries. The dataset is structured into three distinct components: the *Meta Part*, the *Clue Part*, and the *Dialog Part*. We leverage structured geographic knowledge, elicited from expert GeoGuessr players, and human-guided interactions with powerful MLLMs to facilitate the construction of this dataset.

3.1 Meta Part

In the Meta Part, images and meta-geographic information are taken from the OSV-5M [2], which inherits its characteristics of good distribution, wide scope, and high quality. After removing a small number of samples with incomplete location annotations, we organize each sample into JSON format using three levels of administrative boundaries—country, region, and city. This results in a total of 4.87 million entries, covering 70k cities, 2.7k regions, and 210 countries.

3.2 Clue Part

To enhance interpretability, we develop an automated multimodal QA generation paradigm that transforms raw geographic annotations into structured image-text clue pairs. Figure 2(a) illustrates this high-correlation data generation pipeline.

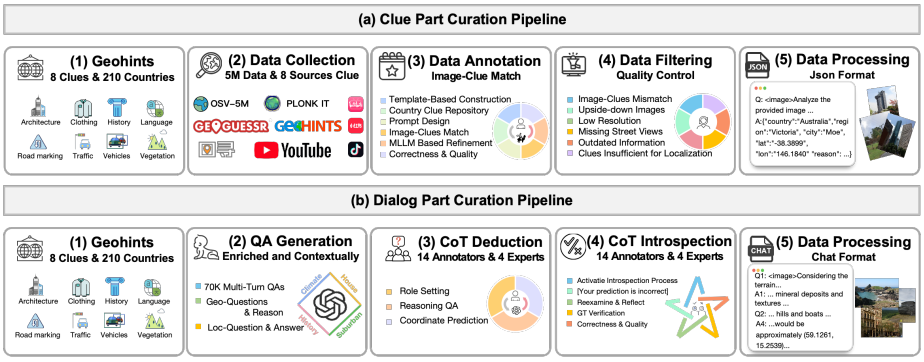

Fig. 2: An illustration of our pipeline for data curation. (a) We construct the Clue Part by leveraging guidance clues from online geolocation game communities and employing an MLLM. (b) We generate location-agnostic, multi-turn reasoning QA pairs and high-quality dialog data for the Dialog Part, applying the Interactive Reasoning CoT method to activate CoT Deduction and CoT Introspection tasks.

MLLM-Based Refinement. While the 3,000 expert clues from GeoGuessr and Tuxun provide rich localization knowledge, pure text lacks the visual context necessary for effective multimodal reasoning. To address this, we leverage MLLMs to align each clue with its corresponding visual features. We sampled 70k globally distributed images from OSV-5M [2] and implemented a two-step process: (1) *constructing country-specific clue repositories* via manual classification, and (2) *automated image-clue matching* to associate street-view imagery with relevant regional descriptors.

Human Verification. Despite rigorous automated matching, visual ambiguities (e.g., low resolution or missing views) may persist. We implemented a manual validation protocol to identify and rectify erroneous pairs. Evaluators trace error sources to either prune problematic samples or recalibrate metadata, ensuring high fidelity between natural language descriptions and their visual targets.

3.3 Dialog Part

As shown in Figure 2(b), we begin the *Dialog Part* construction process by standardizing a well-annotated subset of the Google Landmark V2 into a unified metadata structure, ensuring the generation of multi-turn reasoning QA pairs that are location-agnostic. In order to enhance GaGA's reasoning depth and conversational ability by supporting the analysis of images from multiple perspectives and inferring specific locations, we select 73K samples from Google Landmark V2 with rich information such as architecture, vegetation, cultural elements, and climate. Then, with the assistance of GPT-4V, we generate QA pairs using the Interactive Reasoning CoT method.

Question-Answer Generation We intend to create image descriptions that thoroughly capture visible appearance and attributes, integrating relevant knowledge, climatic characteristics, architectural styles, and even historical context. This all-encompassing strategy ensures the dataset's robust support for a broad spectrum of real-world applications by providing enriched and contextually rich data.

The generation of multi-turn QAs mainly relies on providing unified metadata and carefully designed prompts to MLLMs, specifically GPT-4V. Through this process, GPT-4V engages in multi-turn self-questioning based on the image, gradually guiding the model to reason through and uncover the geographical information. Each set of multi-turn QAs includes the following key attributes: *question ID, source dataset, image path, three geo-questions w/ reasoning process, and one loc-question w/ ultimate answer*. This structure ensures the logical coherence of the multi-turn QAs and clearly presents the progression from question to reasoning process to the final answer.

Prompting techniques improve LLMs' reasoning and problem solving abilities across diverse tasks [17, 20, 31]. We integrate the images' unified metadata format to generate high-quality dialog data. Using the Interactive Reasoning CoT method, we activate two tasks: *CoT Deduction* and *CoT Introspection*. In the next part, we elaborate on the implementation details of these two tasks.

CoT Deduction. To extract the reasoning chain behind the geographic location predictions from GPT-4V as the training data, we explicitly extract the reasoning chain supporting the model's QA process. Specifically, we draw on the concept of interactive reasoning from reinforcement learning and propose the *CoT Deduction* method to handle the geographic location prediction task.

The *CoT Deduction* consists of three parts: *(1) Role Setting*. In CoT Deduction, we set up two roles: *Geo-Guessr player* and *questioner*. The questioner and player interact, with the questioner asking questions and the player responding based on the image clues and existing knowledge. *(2) Reasoning QA*. We aim to explicitly extract the internal principles of geographic location reasoning to construct MG-Geo's Dialog Part. For each question from the questioner, the Geo-Guessr player gradually deduces the geographic location based on various aspects embedded in the image, such as the environment and climate, architecture and landmarks, language and culture, and people's appearance. Each QA round (*i.e.*, Q1A1, Q2A2, and Q3A3) helps the player narrow down the possibilities, gradually approaching the correct answer. *(3) Coordinate Prediction*. After a series of reasoning steps, the player needs to provide a specific geographic coordinate and briefly explain their choice (Q4A4).

After *CoT Deduction* generates the predicted coordinates, we initiate a *Decision Criterion* to evaluate the predicted coordinates' accuracy. Specifically, we calculate the Haversine distance between the predicted coordinates and the unified metadata. If the distance between the predicted and true coordinates is greater than $25km$, the *CoT Introspection* process is triggered.

CoT Introspection. If the initial coordinate prediction deviates from the metadata by more than 25km, we trigger CoT Introspection. Instead of treating

the ground-truth location as a shortcut answer, this stage asks the MLLM to re-check whether each reasoning step is visually grounded, whether the selected clues are geographically discriminative, and whether any high-specificity evidence has been overlooked. The resulting sample is retained only when the revised reasoning remains supported by visible image evidence and passes the quality-control criteria. We explicitly tag all such samples as `introspection`; in MG-Geo, 70.1% of the Dialog Part, approximately 51K/73K samples, follows this path. On the triggered subset, the original pre-refinement deduction already reaches 85.2% Acc@750km and 60.2% Acc@200km, suggesting that Introspection mainly refines directionally correct but insufficiently fine-grained reasoning rather than replacing arbitrary predictions.

4 The GaGA Model

Capitalizing on the MG-Geo dataset, we present GaGA, an interactive MLLM designed to transcend the *"black box"* nature of traditional geolocation. As illustrated in Figure 3, GaGA shifts the geolocation paradigm from static coordinate regression to a dynamic reasoning process. By integrating robust spatial awareness with extensive world knowledge, GaGA can effectively leverage logical constraints during user interaction to refine its predictions in real-time.

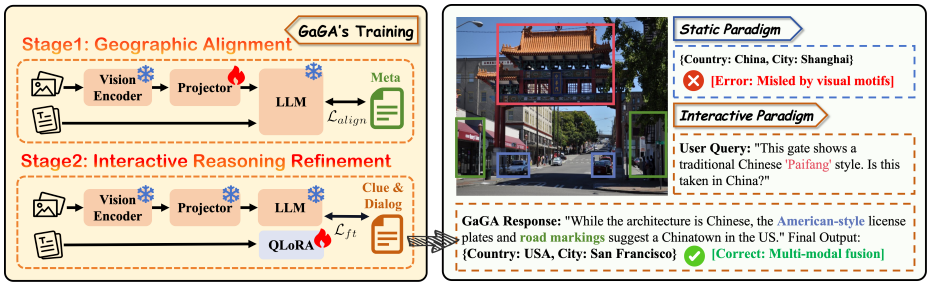

Fig. 3: Overview of the GaGA framework. We transition the geolocation task from a static classification paradigm to a dynamic interactive paradigm, enabling the model to refine predictions through logical constraints. The architecture utilizes a dual-stream training process to bridge the gap between visual motifs and geographic knowledge.

4.1 Model Architecture

GaGA adopts a modular MLLM architecture following the LLaVA paradigm [23], optimized for iterative spatial reasoning. The framework integrates a vision encoder f_V, a cross-modal projector f_P, and an LLM f_L (Llama3-8B [1]) as the central reasoning engine.

Visual Encoding and Projection. For an input street-view image X_v, we employ a pretrained CLIP-ViT [12] as the vision encoder f_V to extract high-level

semantic features. To bridge the modality gap, the projector f_P (a two-layer MLP) maps these visual representations into the LLM's latent word embedding space. The resulting visual tokens are obtained via $E_v = f_P(f_V(X_v))$.

Multimodal Contextualization and Reasoning. To support the interactive geolocation paradigm, the text tokenizer f_T is designed to process not just a single query, but a progressive reasoning trajectory. It transforms the cumulative dialogue history $H_{<n}$ and the current user query Q_n (which may contain high-specificity geographic priors, such as language or traffic rules) into textual tokens $E_{T,n} = f_T([H_{<n}; Q_n])$. These tokens are concatenated with E_v to form a unified input sequence. The LLM then performs auto-regressive inference to generate a refined reasoning trajectory $R_n = f_L([E_v; E_{T,n}])$. This architecture facilitates a deep cross-modal exchange, allowing f_L to contextualize abstract visual patterns within the provided geographic constraints, sequentially updating its belief state to escape visual ambiguities.

4.2 Training Strategy and Objectives

The training protocol of GaGA is conceptualized as a two-stage paradigm: *Geographic Alignment* and *Interactive Instruction Tuning*. This progressive approach ensures that the model first internalizes a global geographic world-view before mastering complex spatial reasoning and dialogue capabilities.

Stage 1: Geographic Alignment. The primary objective of this stage is to bridge the modality gap between generic visual features and structured geographic metadata. We initialize the multimodal projector with ShareGPT4V [5] weights to leverage its robust cross-modal mapping. During this phase, both the vision encoder and the LLM backbone are kept frozen to preserve their respective pre-trained representational power.

We optimize the projector ϕ by minimizing the standard cross-entropy loss over the *Meta Part* of the MG-Geo dataset. Formally, given an image X_v and its associated geographic metadata Y_{geo} (comprising country, region, city, and coordinates), the training objective is:

$$\mathcal{L}_{align} = -\sum_{t=1}^{T} \log P(y_t | X_v, y_{<t}; \phi), \tag{1}$$

where $Y_{geo} = \{y_1, ..., y_T\}$ represents tokenized geographic attributes. This stage forces the projector to learn a *Spatial-Aware Embedding* that aligns visual patterns with specific global hierarchies.

Stage 2: Interactive Reasoning Refinement. The second stage focuses on cultivating **Constraint-based Reasoning** and multi-turn adaptability. We freeze the projector ϕ and apply QLoRA [11] to the LLM to optimize its decision-making logic. The training corpus integrates the Meta, Clue, and Dialog parts of MG-Geo, with a strategic emphasis on the *Interactive Reasoning CoT* strategy. Crucially, the training corpus here is a heterogeneous mixture of 240k carefully curated samples from the Meta, Clue, and Dialog parts of MG-Geo. The *Clue Part* explicitly exposes the model to 8 distinct dimensions of geographic evidence

(e.g., architecture, vegetation, road markings), while the *Dialog Part* instills the *Interactive Reasoning CoT* strategy. The optimization goal is to maximize the conditional likelihood of the correct reasoning chains R and the final coordinates C given the visual input X_v and user queries Q:

$$\mathcal{L}_{ft} = -\mathbb{E}_{(X_v,Q,R,C)\sim\mathcal{D}_{MG-Geo}}[\log P(R, C|X_v, Q; \theta_{LoRA})]. \tag{2}$$

By tightly coupling the LLM optimization with MG-Geo's conversational structure, this multi-task objective enables GaGA to identify and rectify its initial misjudgments. This ensures that the final response is the result of a deliberate, multi-turn evidentiary synthesis rather than a simplistic statistical mapping.

5 Experiment

To demonstrate the efficacy of our dataset in addressing the geographic knowledge gap in existing models and to showcase GaGA's ability in geolocation tasks, we conducted a comprehensive suite of experiments, the primary findings of which are presented in this section.

5.1 Experimental Setup

Metrics. We use the following metrics to evaluate the prediction accuracy of the geolocation model:

(1) Accuracy of predicted geographical names at various administrative levels: country, region and city.

(2) Accuracy of predicted coordinates within various distance thresholds: 1km, 25km, 200km, 750km, and 2500km, calculated as the haversine distance between the model's predicted GPS coordinates and the ground truth.

(3) Geoscore: it is defined as $5000exp(-\delta/1492.7)$ based on the famous "Geoguessr" game. δ represents the Haversine distance between predicted and ground truth image locations.

Benchmark. We assess GaGA's performance across multiple benchmarks. For its primary global geolocation capability, we use GWS15k [8] and OSV-5M-test [2] for its balanced worldwide coverage. Since GWS15k is not open-source, we reproduce it in this study. The pseudocode and data distribution are presented in the Appendix. To further evaluate its generalization capability, we conduct comparative experiments between GaGA and state-of-the-art methods on two web-crawled image benchmarks, Im2GPS3k [35] and YFCC4k [33], with results reported in the Appendix.

Implementation Details. All experiments are conducted using the XTuner platform [10], facilitating efficient multimodal model tuning and deployment. For reasoning tasks, we use LMDeploy [9] tool. We conduct all the experiments on 8 $\times$ RTX4090 GPUs.

Table 2: Administrative-Level Accuracy and Coordinates Accuracy of GaGA on GWS15K Bench. Left: Administrative-Level. **Right:** Coordinates Accuracy. We use **bold** to indicate the best performance, and '___' for the second-best, respectively. ★ represents the model evaluated on GWS15k reproduced in this paper.

Method	Recall	Admin-Level(%)	
		Country	City
StreetCLIP [15]	1	40.11	3.02
Hybrid [2]	1	**58.49**	**3.36**
LLaVA-Llama3	0.99	1.76	0.02
InternVL2-8B [6]	0.96	24.74	0.48
Qwen2.5-VL-7B [4]	0.99	33.37	1.43
Qwen3-VL-8B [3]	0.99	39.10	1.63
GeoReasoner [21]	1	40.63	1.11
GLOBE [22]	1	47.09	3.44
GaGA(Ours)	1	**63.06**	**6.28**

Method	Coordinates Accuracy (% @ km)					Geoscore
	1km	25km	200km	750km	2500km	
ISNs [24]	0.05	0.6	4.2	15.5	38.5	-
Translocator [25]	0.5	1.1	8	25.5	48.3	-
GeoDecoder [26]	**0.7**	1.5	8.7	26.9	50.5	-
GeoCLIP★ [34]	0.2	3.1	15.4	40.3	71.2	2345.2
PIGEOTTO [16]	**0.7**	9.2	31.2	**65.7**	**85.1**	-
Hybrid★ [2]	0.08	**14.9**	**39.3**	56.2	74.4	**2944.9**
GeoReasoner★ [21]	0.04	2.1	9.8	32.9	64.0	2018.4
GRE [36]	**0.9**	4.1	18.9	54.8	78.3	-
GLOBE★ [22]	0.1	3.9	15.0	40.8	69.4	2311.3
GaGA★(Ours)	0.1	**8.5**	**33.9**	**60.6**	**82.2**	**3113.0**

5.2 Geolocation Performance

The evaluation results on the GWS15k benchmark are summarized in Table 2, where we categorize current methodologies into traditional discriminative models and MLLM-based reasoning assistants.

As shown in Table 2 (Left), GaGA achieves outstanding performance in administrative boundary identification. Notably, while general-purpose MLLMs (e.g., LLaVA-Llama3, Qwen-VL [3, 4]) and even specialized geographic MLLMs (e.g., GeoReasoner [21], GLOBE [22]) struggle with precise localization, GaGA achieves SOTA performance among all MLLM-based methods. Specifically, GaGA attains a country-level accuracy of 63.06% and a city-level accuracy of 6.28%, outperforming the previous best-performing MLLM, GLOBE, by a substantial margin of 15.97% and 2.84% respectively. Even when compared to traditional SOTA classifiers like Hybrid [2], GaGA maintains a clear lead (+4.57% at country level and +2.92% at city level), demonstrating that the interactive reasoning paradigm does not sacrifice, but rather enhances, fundamental localization precision. Furthermore, GaGA maintains a 100% Recall rate, ensuring that every query is met with a valid geographic coordinate—a critical reliability metric where other LLMs often falter.

Table 2 (Right) presents the coordinate-level error distribution. Among MLLM-based approaches (highlighted in the bottom section), GaGA consistently defines the state-of-the-art, significantly surpassing GeoReasoner, GRE, and GLOBE across thresholds of 25 to 2500km. For instance, at the 200km threshold—a key metric for regional localization—GaGA leads the second-best MLLM (GRE) by 15.0%. When compared to traditional end-to-end models, GaGA exhibits competitive performance, securing the second-best results from 200km to 2500km. Most importantly, GaGA achieves the **highest Geoscore (3113.0)** among all reproducible models, including the strong Hybrid baseline. This peak Geoscore underscores GaGA's unique ability to strike an optimal balance between high-

Table 3: Performance of GaGA with Tiered Interactive Guidance.

Interaction Tiers	Admin-Level Accuracy (%)		
	Country	Region	City
Direct Inquiry	64.89	27.97	7.67
Focus-based Inquiry	61.24	29.25	8.22
	−3.65	+1.28	+0.55
Knowledge-augmented Guidance	**74.77**	**34.73**	**9.87**
	+9.88	+6.76	+2.20

Table 4: Impact of Clue Ambiguity on Geolocation Performance.

Type of Subset	Admin-Level Accuracy (%)		
	Country	Region	City
Ambiguous	59.5	26.2	7.0
High-Specificity	**62.2**	**28.9**	**9.1**

precision hits and the mitigation of catastrophic outliers, a direct result of its clue-based reasoning and interactive refinement capabilities.

Table 6 reports the comparative results on the OSV-5M test set. Notably, GaGA achieves the highest City-level accuracy (7.4%) among all competing methods. In terms of global metrics, GaGA reaches a Geoscore of 3413 and a Country-level accuracy of 71.4%, consistently outperforming specialized models like ISNs and the Hybrid baseline. These findings underscore GaGA's robustness in balancing macro-scale region recognition and micro-scale city identification.

5.3 Inference-Time Interactive Geolocation Analysis

A fundamental strength of GaGA lies in its capacity for **Multi-tier Interactive Reasoning**, allowing it to dynamically refine geolocation beliefs through dialogue. In this section, we provide a formal dissection of this capability. First, we establish a taxonomic framework for interactive guidance to quantify performance gains across different levels of information density. Second, we investigate the "Similarity Trap" to elucidate how the specificity of geographic cues influences the model's posterior distribution.

Effectiveness of Tiered Interactive Guidance To evaluate GaGA's reasoning flexibility, we curated a diagnostic set of 547 images spanning diverse cultural and natural landscapes. We standardized the interaction into three progressive tiers: (1) **Zero-shot (Direct Inquiry)**: Baseline prediction without external cues. (2) **Tier-1 (Focus-based Inquiry)**: Providing a guiding question (e.g., *"Considering the architectural design, what region of the world would you think displays such forms?"*) to direct the model's attention to specific attributes without providing factual answers. (3) **Tier-2 (Knowledge-augmented Guidance)**: Providing both a question and a factual geographic prior (e.g., *"The photo shows the typical oceanic and continental climate zones; which countries typically use this color?"*).

Table 3 illustrates the performance dynamics. The results reveal several key insights. Firstly, **Tier-2 interaction** yields the most significant gains, with GaGA's country-level accuracy surging by 9.88% (reaching 74.77%). This demonstrates GaGA's superior ability in *Knowledge Integration*—effectively using external anchors to prune the search space. Secondly, an intriguing observation arises in **Tier-1 interaction**: while region and city-level precisions improve, country-level

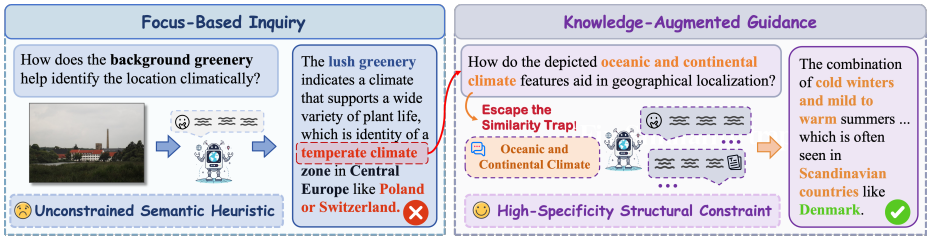

Fig. 4: Navigating through the "Similarity Trap" via tiered interaction.

accuracy experiences a slight decline (-3.65%). We characterize this as the exploration cost within the *Similarity Trap*. Without a factual anchor, Tier-1 queries compel the model to focus on broadly distributed visual features. While this sharpens the model's focus on local details (improving city-level hits), it simultaneously increases the risk of being misled by cross-border visual similarities.

Geographic Ambiguity and the "Similarity Trap" Global geolocation is inherently plagued by visual ambiguity, a challenge we formalize as the **"Similarity Trap"**. This occurs when geographically widespread features—such as neoclassical architecture or generic temperate flora—produce overlapping representations in the model's feature space, leading to over-confident misclassifications [13]. While static retrieval-based models are often paralyzed by such ambiguity, GaGA's interactive framework introduces a dynamic *Error-Correction Mechanism*.

As illustrated in the comparative analysis in Figure 4, when GaGA is prompted with an **Unconstrained Semantic Heuristic** (e.g., *"How does the background greenery help identify the location?"*), it may initially succumb to the trap by over-relying on generic motifs that lack geographic exclusivity. However, by transitioning to a **High-Specificity Structural Constraint** (e.g., *"How do the oceanic and continental climate features aid in localization?"*), the model is forced to decouple ambiguous visual signals and engage in constraint-based reasoning. This logical anchoring allows GaGA to suppress noise and re-route its reasoning toward the correct manifold.

Formally, let G denote the geographic label and c a visual clue. We define clue specificity as

$$s(c) = 1 - \frac{H(G|c)}{H(G)},$$

where low-specificity clues have high conditional entropy over possible locations, while high-specificity clues sharply reduce the uncertainty. Given model salience weights α_i over visual clues $\{c_i\}$, a Similarity Trap occurs when the prediction is dominated by low-specificity clues, i.e., $\sum_i \alpha_i(1 - s(c_i))$ is high, leading to a localization error $d(\hat{y}, y^*) > \tau$.

Based on this specificity criterion, we partitioned our evaluation set into two diagnostic subsets based on the information entropy and geographic exclusivity of the cues:

(1) High-Specificity Subset: Focuses on *Deterministic Markers* with narrow geographic distributions and strong logical constraints (*e.g.*, road markings, signage language, or driving side).

(2) Ambiguous Subset: Focuses on *Distributive Features* with high spatial variance and weak exclusionary power (*e.g.*, architectural styles, general vegetation, or terrain).

The empirical results in Table 4 confirm that high-specificity cues yield significantly higher precision across all administrative levels. This validates our hypothesis that the "Similarity Trap" is the primary bottleneck in geolocation. More importantly, it underscores the necessity of GaGA's interactive paradigm: by converting human-provided high-specificity priors into logical constraints, GaGA effectively navigates through visual ambiguity where static models fail.

5.4 Ablation Experiments

To systematically evaluate the contribution of each component within the MG-Geo dataset, we conduct a comprehensive ablation study. Using the test set from Section 5.3, we design five experimental groups. We evaluate performance across four key metrics: *Accuracy* (Geo-Score), *Informativeness* (hit rate of generated clues, assessed by an LLM Judge), *Fluency* (Likert scale score for coherence, assessed by an LLM Judge), and *Relevance* (contextual relevance, assessed by an MLLM Judge).

As detailed in Table 5, our ablation study validates the indispensable roles and intentional design trade-offs of the MG-Geo dataset components. The `Meta` data serves as the bedrock for static localization, achieving the highest overall Accuracy (3872). Conversely, the `Dialog` component acts as the primary driver for conversational reasoning, yielding substantial improvements across interactive metrics, including a significant leap in Informativeness (from 0.6164 to 0.7549). Furthermore, the `Clue` data functions as a conditional enhancer; while it can slightly perturb accuracy when lacking conversational context, combining it with the reasoning framework (`Meta`+`Clue`+`Dialog`) maximizes both Informativeness (0.7809) and Fluency (4.9555). Ultimately, this deliberate trade-off—sacrificing a marginal degree of raw precision for massive gains in explainability and interaction—confirms our core objective of developing a balanced, interactive global geolocation assistant.

We further isolate the effect of CoT Introspection by removing all Introspection-tagged dialogs while maintaining the same Stage-2 sample budget. The full model improves over the w/o-Introspection variant by +5.74 Acc@200km, +12.77 Acc@750km, +15.00 Acc@2500km, and +7.23 country-level accuracy, supporting the role of Introspection as a cue-correction signal rather than a coordinate shortcut. Detailed results are provided in the Appendix.

5.5 Release, Privacy, and Responsible Use

MG-Geo is released in a source-compatible form, including annotations, prompts, splits, generation-path tags, and reconstruction scripts.

Table 5: Ablation study on the components of the MG-Geo dataset.

Training Data	Acc.	Info.	Flu.	Rel.
Meta	**3872**	0.616	4.572	2.294
Meta+Clue	3835	0.617	4.494	2.168
Meta+Dialog	3854	0.754	4.929	3.579
Meta+Clue+Dialog	3817	**0.780**	**4.955**	3.650
Clue+Dialog	3773	0.767	4.925	**3.898**

Table 6: Performance comparison on OSV-5M-test set.

Model	Geoscore	Admin-Level	
		Country	City
ISNs [24]	3331	66.8	4.2
Hybrid [2]	3361	67.4	<u>6.0</u>
RFM $\mathcal{S}_2$ [13]	**3767**	**76.2**	5.4
GaGA(Ours)	<u>3413</u>	<u>71.4</u>	**7.4**

6 Conclusion

In this work, we address key challenges in global geolocation by tackling the critical lack of comprehensive data and the performance limitations of existing methods. To this end, we introduce **MG-Geo**, the first large-scale, high-quality multimodal dataset specifically designed with rich geographic cues to bridge this knowledge gap for MLLMs. Building upon this foundational dataset, we developed **GaGA**, a novel MLLM that not only demonstrates superior performance over existing models in predicting administrative boundaries but also introduces a crucial interactive capability for refining localization through user dialogue. Our work underscores the critical role of high-quality, domain-specific datasets in unlocking the potential of MLLMs for complex spatial reasoning tasks and opens new avenues for a wide range of geographic downstream applications.

Acknowledgements

This work was supported in part by the Key Deployment Program of the Chinese Academy of Sciences, China under Grant KGFZD-145-25-39, the National Natural Science Foundation of China under Grants 62272438, and Beijing Natural Science Foundation L25700.

References

1. AI@Meta: Llama 3 model card. None (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
2. Astruc, G., Dufour, N., Siglidis, I., Aronssohn, C., Bouia, N., Fu, S., Loiseau, R., Nguyen, V.N., Raude, C., Vincent, E., et al.: Openstreetview-5m: The many roads to global visual geolocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21967–21977 (2024)
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions (2023), https://arxiv.org/abs/2311.12793
6. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821 (2024)
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023)
8. Clark, B., Kerrigan, A., Kulkarni, P.P., Cepeda, V.V., Shah, M.: 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. pp. 23182–23190 (2023)
9. Contributors, L.: Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy (2023)
10. Contributors, X.: Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner (2023)
11. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems **36** (2024)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
13. Dufour, N., Kalogeiton, V., Picard, D., Landrieu, L.: Around the world in 80 timesteps: A generative approach to global visual geolocation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23016–23026 (2025)
14. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M.,

Krishna, R., Koh, P.W., Saukh, O., Ratner, A., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., Schmidt, L.: DataComp: In search of the next generation of multimodal datasets. NeurIPS Dataset and Benchmark (2023)
15. Haas, L., Alberti, S., Skreta, M.: Learning generalized zero-shot learners for open-domain image geolocalization. arXiv preprint arXiv:2302.00275 (2023)
16. Haas, L., Skreta, M., Alberti, S., Finn, C.: Pigeon: Predicting image geolocations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12893–12902 (2024)
17. Huang, W., Abbeel, P., Pathak, D., Mordatch, I.: Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In: International conference on machine learning. pp. 9118–9147. PMLR (2022)
18. Jia, P., Liu, Y., Li, X., Zhao, X., Wang, Y., Du, Y., Han, X., Wei, X., Wang, S., Yin, D.: G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. Advances in Neural Information Processing Systems **37**, 53198–53221 (2024)
19. Jia, P., Park, S., Gao, S., Zhao, X., Li, Y.: Georanker: Distance-aware ranking for worldwide image geolocalization. arXiv preprint arXiv:2505.13731 (2025)
20. Jimenez, C.E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., Narasimhan, K.: Swe-bench: Can language models resolve real-world github issues?, 2024. URL https://arxiv. org/abs/2310.06770 (2023)
21. Li, L., Ye, Y., Jiang, B., Zeng, W.: Georeasoner: Geo-localization with reasoning in street views using a large vision-language model (2024)
22. Li, L., Zhou, Y., Liang, Y., Tsung, F., Wei, J.: Recognition through reasoning: Reinforcing image geo-localization with large vision-language models. In: Advances in Neural Information Processing Systems (2025)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), https://arxiv.org/abs/2304.08485
24. Muller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation estimation of photos using a hierarchical model and scene classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 563–579 (2018)
25. Pramanick, S., Nowara, E.M., Gleason, J., Castillo, C.D., Chellappa, R.: Where in the World is this Image? Transformer-based Geo-localization in the Wild. In: Proceedings of the European Conference on Computer Vision. pp. 196–215 (2022)
26. Qi, F., Dai, M., Zheng, Z., Wang, C.: Geodecoder: Empowering multimodal map understanding. arXiv preprint arXiv:2401.15118 (2024)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
28. Roberts, J., Lüddecke, T., Das, S., Han, K., Albanie, S.: Gpt4geo: How a language model sees the world's geography. arXiv preprint arXiv:2306.00020 (2023)
29. Seo, P.H., Weyand, T., Sim, J., Han, B.: CPlaNet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps. In: Proceedings of the European Conference on Computer Vision. pp. 536–551 (2018)
30. Singh, S., Fore, M., Stamoulis, D.: Geollm-engine: A realistic environment for building geospatial copilots. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 585–594 (2024)
31. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11888–11898 (2023)

32. Theiner, J., Müller-Budack, E., Ewerth, R.: Interpretable semantic photo geolocation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 750–760 (2022)
33. Thomee, B., Elizalde, B., Shamma, D.A., Ni, K., Friedland, G., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the Acm **59**(2), 64–73 (2016)
34. Vivanco Cepeda, V., Nayak, G.K., Shah, M.: Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. Advances in Neural Information Processing Systems **36** (2024)
35. Vo, N., Jacobs, N., Hays, J.: Revisiting im2gps in the deep learning era. In: Proceedings of the IEEE international conference on computer vision. pp. 2621–2630 (2017)
36. Wang, C., Ye, X., Pan, X., Pan, Z., Wang, H., Song, Y.: Gre suite: Geo-localization inference via fine-tuned vision-language models and enhanced reasoning chains (2025)
37. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2575–2584 (2020)
38. Weyand, T., Kostrikov, I., Philbin, J.: Planet: Photo geolocation with convolutional neural networks. In: European Conference on Computer Vision. pp. 37–55 (2016)
39. Zhang, X., Li, X., Sultani, W., Zhou, Y., Wshah, S.: Cross-view geo-localization via learning disentangled geometric layout correspondence. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3480–3488 (2023)
40. Zhang, X., Sultani, W., Wshah, S.: Cross-view image sequence geo-localization. In: WACV (2023)
41. Zhou, F., Qi, X., Zhang, K., Trajcevski, G., Zhong, T.: Metageo: A general framework for social user geolocation identification with few-shot learning. IEEE Transactions on Neural Networks and Learning Systems **34**(11), 8950–8964 (2023). https://doi.org/10.1109/TNNLS.2022.3154204
42. Zhou, Z., Zhang, J., Guan, Z., Hu, M., Lao, N., Mu, L., Li, S., Mai, G.: 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. pp. 2749–2754 (2024)
43. Zhu, S., Shah, M., Chen, C.: TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1162–1171 (2022)