GeoBayes: Probabilistic Image Geo-Localization Inference via Sequential Bayesian Updating

Weimin Shi[1,2], Xiang Li[1], Kaige Li[3], Junhao Fang[1], Qiang Zhou[1], Qichuan Geng[4*], Zhong Zhou[1, 2*]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China
[3]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[4]The Information Engineering College, Capital Normal University, Beijing, China
{shiwm, lx_7342, nosta1gia, zhouq, zz}@buaa.edu.cn, likg@mail.sysu.edu.cn, gengqichuan1989@cnu.edu.cn

Abstract

Image geo-localization aims to determine the geographic location of a query image. While Multimodal Large Language Models (MLLMs) show potential for this task due to their rich world knowledge and explainability, they often struggle with confirmation bias, i.e., committing prematurely to potentially incorrect guesses driven by visual clues with diverse geographic likelihoods. In this paper, we propose GeoBayes, a novel training-free framework that formulates geo-localization as a Maximum a Posteriori (MAP) estimation task over multiple geographic hypotheses and performs probabilistic reasoning via sequential Bayesian updating. GeoBayes regards each visual object and its associated geographic clues as probabilistic evidence, integrating them iteratively through a Hypothesize–Verify–Update loop. At each step, it evaluates how new evidence supports existing hypotheses and updates their posterior probabilities, gradually converging on the most probable location. This allows GeoBayes to explicitly quantify and fuse the varied geographic probabilities implied by diverse visual elements, reducing the risk of overcommitting to misleading clues. Furthermore, considering the natural hierarchy of geographic labels (e.g., country, city), GeoBayes introduces a state memory mechanism that stores hypotheses, inference context, and evidence scores across levels. This design enables the framework to propagate prior knowledge across levels of the geographic hierarchy and incorporate geographic structural constraints into the Bayesian update process, achieving a coarse-to-fine geo-localization. Experiments on IM2GPS3k and YFCC4K show that GeoBayes improves MLLM-based geo-localization accuracy without extra training. This demonstrates the effectiveness of probabilistic reasoning for robust and interpretable geo-localization.

Introduction

Image geo-localization aims to pinpoint the global location of a query image by associating visual elements with geographic clues. It has broad applications in smart cities (Mandal 2024), intelligence analysis (Wang et al. 2024), and public safety monitoring (Hu et al. 2024). To address this, existing methods typically fall into three types: classification-, retrieval-, and generation-based methods. Classification methods (Seo et al. 2018; Theiner, Muller-Budack, and Ewerth 2022; Haas et al. 2024) divide the Earth's surface into predefined grids and predict coordinates by identifying the corresponding grids. Retrieval methods (Zhu et al. 2023; Vivanco Cepeda, Nayak, and Shah 2023; Zemene et al. 2018) match similar samples in large-scale geolabeled databases and return their GPS coordinates. Both types rely on timeliness and data coverage, limiting their effectiveness (Li et al. 2024; Manvi et al. 2023).

Recently, multimodal large language models (MLLMs) have demonstrated powerful world knowledge and scene understanding capabilities, enabling deep semantic parsing of visual input (Jiang et al. 2024; Cheng et al. 2024). Building on this advancement, generation-based methods (Jia et al. 2024; Zhou et al. 2024; Han et al. 2024) leverage MLLMs to directly reason about geographic clues, enabling more interpretable and flexible localization. Further, since MLLMs are general-purpose models, some methods fine-tune them for

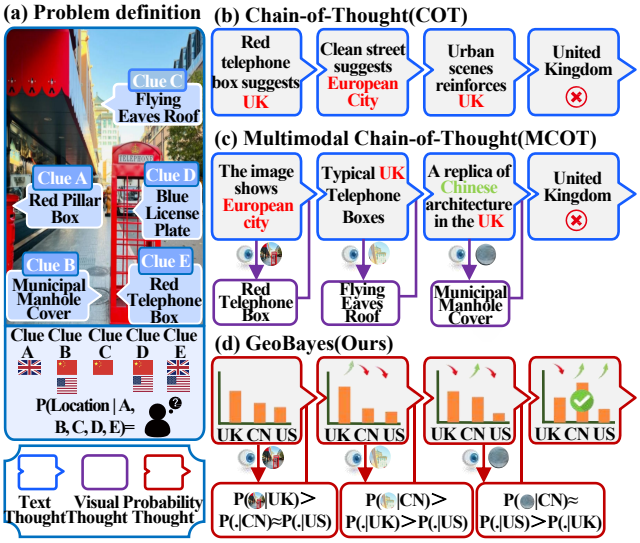

Figure 1: Comparison of reasoning paradigms. CoT and MCoT are prone to confirmation bias, while GeoBayes mitigates this by progressively updating geographic hypotheses via the proposed Probability Thought paradigm.

*Corresponding authors.

geo-localization tasks (Li et al. 2024; Zhang et al. 2025a,b). However, generation-based methods often struggle to accurately quantify different geographical possibilities associated with visual elements. As shown in Fig. 1(a), a British postbox may appear in both the UK and the US, but is more common in the UK, whereas flying-eaved roofs are highly indicative of China. Yet, MLLMs' chain-of-thought (CoT) reasoning fails to accommodate such varied geographic priors, often reinforcing the initial hypothesis (e.g., UK) without reconsidering alternatives as new evidence emerges. For example, in Fig. 1(b), the model locks onto a European city based on the telephone box and overlooks contradictory cues such as manhole covers. Similarly, in Fig. 1(c), even when multimodal information is incorporated, the reasoning still orients around the initial hypothesis, which suffers from confirmation bias. As a result, early errors accumulate during the reasoning without correction.

To this end, we propose GeoBayes, a training-free framework that casts geo-localization as a Maximum a Posteriori (MAP) estimation task solved by sequential Bayesian Updating. GeoBayes introduces a "Probability Thought" paradigm, enabling MLLMs to iteratively refine predictions by weighing evidence over multiple geographic hypotheses instead of committing to early guesses. Specifically, GeoBayes operates via a Hypothesize–Verify–Update loop, generating candidate locations and their prior probabilities. It then verifies new visual objects, either via internal reasoning or external retrieval, and converts them into structured evidence. A comparative scoring function quantifies how strongly each piece of evidence supports each hypothesis, forming likelihood estimates. These likelihoods are then fused with the prior via a sequential Bayesian update, gradually refining the hypothesis distribution and steering the model toward the most probable location. This design mitigates early confirmation bias and refines predictions with accumulated evidence, as shown in Fig. 1(d).

By framing the reasoning process as Bayesian updates over evidence, GeoBayes empowers MLLM to "think probabilistically" without extra training, effectively handling uncertainty and resolving conflicting clues. Extensive experiments on the IM2GPS3k (Thomee et al. 2016) and YFCC4K (Thomee et al. 2016) datasets show that GeoBayes improves localization accuracy by 3.2% on average compared with the state-of-the-art(SOTA) generation-based methods. Our main contributions are as follows:

- We formalize image geo-localization as Bayesian MAP inference, using probability reasoning to quantify and combine diverse geographic clues, enabling robust geographic reasoning under visual uncertainty.
- We propose GeoBayes, a hierarchical "Hypothesize–Verify–Update" framework that performs probabilistic inference by iteratively integrating evidence and refining hypotheses, effectively reducing confirmation bias and improving accuracy.
- Extensive experiments show that GeoBayes enhances the geographic reasoning ability of MLLMs without additional training, achieving SOTA performance among generation-based methods.

Related Work

Image Geo-localization. Existing methods fall into three types: 1) Classification-based methods convert coordinate regression into classification by partitioning the Earth into discrete cells, with improvements focusing on partition strategies (Weyand, Kostrikov, and Philbin 2016; Theiner, Muller-Budack, and Ewerth 2022; Haas et al. 2024) and network design (Seo et al. 2018; Vo, Jacobs, and Hays 2017; Muller-Budack, Pustu-Iren, and Ewerth 2018; Pramanick et al. 2022; Clark et al. 2023). 2) Retrieval-based methods match query images against geo-tagged databases, emphasizing geographically discriminative features (Zhu et al. 2023; Vivanco Cepeda, Nayak, and Shah 2023; Zemene et al. 2018; Xu et al. 2024) and robustness to cross-view variations (Liu and Li 2019; Zhu, Yang, and Chen 2021; Yang, Lu, and Zhu 2021; Zhu, Shah, and Chen 2022; Shi et al. 2020). However, both rely on costly large-scale datasets that are hard to maintain and scale. 3) Generation-based methods use MLLMs to infer locations by aligning visual features with geographic descriptions (Jia et al. 2024; Li et al. 2024; Han et al. 2024). To enhance performance, retrieval-augmented methods integrate external information (Zhou et al. 2024) and recent works leverage specific visual objects as aids to extract fine-grained details (Zhang et al. 2025a).

Multimodal Reasoning. Multimodal reasoning enhances geo-localization by integrating visual and textual information. A common strategy is to extract visual features and apply text-based CoT reasoning to derive conclusions. However, recent studies reveal that purely textual CoT often lacks the contextual grounding needed for visual tasks. To address this, recent methods convert visual inputs into structured forms such as textual descriptions (Wu et al. 2024), scene graphs (Mitra et al. 2024), or sets of bounding boxes (Corbière et al. 2025). Beyond this, some works directly inject visual features into the reasoning process (Gao et al. 2025), fine-tune models to generate image-grounded CoT (Li et al. 2025), or apply test-time scaling to improve the model's adaptability (Wang et al. 2025). Despite progress, most still rely on linear CoT to integrate information, which may prematurely commit to incorrect judgments due to semantic ambiguity in visual features.

Preliminary

This section presents the basic principles of Bayesian updating, which is a statistical inference method for updating the probability of a hypothesis H based on evidence E, governed by Bayes' theorem:

$$P(H \mid E) \propto P(E \mid H) \cdot P(H), \tag{1}$$

here $P(H \mid E)$ is the posterior probability that H is true after observing E. $P(E \mid H)$ is the likelihood that describes the probability of observing E given H. $P(H)$ is the prior probability, representing the initial probability estimate for H. When evidence arrives sequentially $E_{1:t} = \{e_1, \ldots, e_t\}$, the posterior is updated iteratively:

$$P(H \mid E_{1:t}) \propto P(e_t \mid H, E_{1:t-1}) \cdot P(H \mid E_{1:t-1}), \tag{2}$$

with the previous posterior $P(H \mid E_{1:t-1})$ serving as the new prior. This enables progressive refinement of the hypothesis as more evidence is observed.

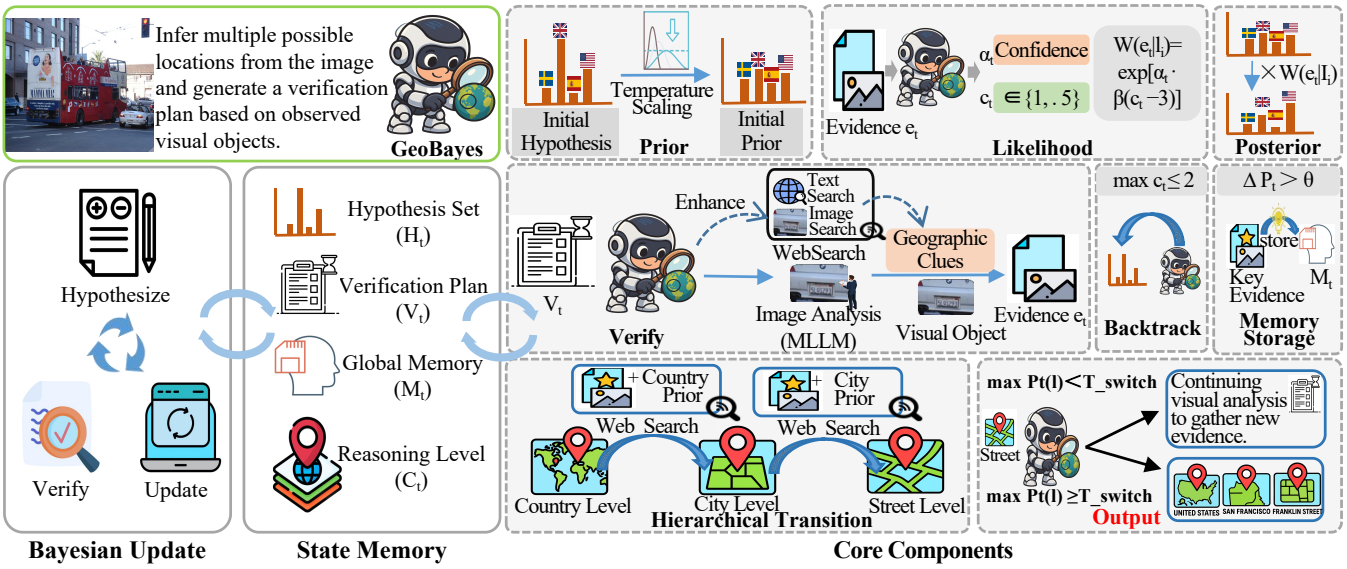

Figure 2: Overview of the GeoBayes framework. GeoBayes combines priors, likelihood, Bayesian updates, evidence collection and storage, and hierarchical transitions to refine location predictions.

Methodology

In this section, we first formulate the geo-localization task as a problem of sequential Bayesian updating and describe how to convert MLLM outputs into prior probabilities and evidence-conditioned likelihoods. We then introduce the state memory mechanism, which maintains hypotheses, evidence, and reasoning context to support long-chain, multi-level inference. Finally, we describe how GeoBayes iteratively updates posteriors via a "Hypothesize–Verify–Update" loop to progressively refine the location estimate. An overview of GeoBayes is shown in Fig. 2.

Task Formulation

Visual elements often imply varied geographic possibilities, e.g., "Gothic architecture" may indicate Germany, France, or the UK with different possibilities. To model such varied associations, we formulate geo-localization as a process of Bayesian sequential inference, which allows the model to reason probabilistically on competing location hypotheses.

Let the hypothesis be "the true geographic location is l", and let each retrieved visual object serve as evidence e_t. The posterior over candidate locations set L is updated iteratively:

$$P(L = l \mid E_{1:t}) \propto P_0(l) \cdot \prod_{t=1}^{n} P(e_t \mid L = l, E_{1:t-1}), \quad (3)$$

where $P_0(l)$ is the prior probability of a location l, and the likelihood $P(e_t \mid L = l, E_{1:t-1})$ captures the probability of observing evidence e_t given location l and past evidence $E_{1:t-1}$. The final estimate l^* is obtained via Maximum a Posteriori (MAP) inference:

$$l^* = \arg\max_{l \in \mathcal{L}} P(L = l \mid E) \quad (4)$$

This step-wise inference gradually rules out unlikely hypotheses, filtering out noise by weighing clues and converging to a location consistent with aggregated evidence.

Prior Probability. To enable probabilistic reasoning, we need to align the outputs of the MLLMs with Bayesian updating. Inspired by recent studies (Han, Buntine, and Shareghi 2024), we instruct MLLM to generate candidate locations l_i with confidence scores s_i from global image analysis. We treat these as the initial prior:

$$P_0(L = l_i) = \frac{\exp(min(s_i, \tau_p)/T)}{\sum_{j=1}^{k} \exp(min(s_j, \tau_p)/T)}, \quad (5)$$

where we apply temperature scaling ($T = 1.5$) and truncate s_i at a maximum of τ_p to smooth overconfident predictions and normalize the prior distribution.

Likelihood Estimation. With the prior probabilities, we need to compute the likelihood $P(e_t \mid l_i)$ for new evidence e_t under each candidate hypothesis l_i. Directly calculating this probability is not feasible (e.g., the probability of finding a white-domed building in the USA). To address this, we design a surrogate scoring function, $W(e_t \mid l_i)$, to approximate the true likelihood.

Specifically, instead of asking the MLLM to output an exact number, we leverage its comparative reasoning ability to evaluate the relevance between the evidence e_t and each candidate hypothesis l_i. For each e_t, the model outputs a support rating $c_t \in \{1, \cdots, 5\}$, where 1 indicates strong contradiction and 5 indicates strong support, along with an associated confidence $\alpha_t \in [0, 1]$, which is then converted to a quantitative evidence support score $W(e_t \mid l_i)$:

$$W(e_t \mid l_i) = \exp[\alpha_t \cdot \beta(c_t - 3)], \quad (6)$$

with $\beta = \ln 2$. This yields symmetric support scores centered at 1, allowing evidence of equal strength but opposite polarity to cancel each other out during Bayesian updates.

We then use W to replace the likelihood term in Eq. 3, enabling the sequential Bayesian update:

$$P(L = l \mid E_{1:t}) \propto P_0(l) \cdot \prod_{t=1}^{n} W(e_t \mid L = l, E_{1:t-1}), \quad (7)$$

here, $W(e_t \mid L = l, E_{1:t-1})$ preserves ranking consistency with the true likelihood, MAP inference remains valid.

State Memory

To support long-chain reasoning, we introduce a state memory mechanism that tracks hypotheses, evidence, memory, and reasoning context throughout the inference process. This design enables hierarchical prior propagation across levels (e.g., country to city) and mitigates the disruptive impact of later ambiguous clues.

State Structure. At each step t during inference, the state memory S_t is composed of four key components:

$$S_t = \{H_t,\ V_t,\ M_t,\ C_t\}, \quad (8)$$

where the **Hypothesis Set** H_t contains all current candidate locations $l_i \in L$ and their corresponding posterior probabilities $P(l_i \mid E_{1:t})$. For instance, H_t could be the distribution $\{\text{United Kingdom} = 0.261,\ \text{United States} = 0.188,\ \text{Sweden} = 0.185\}$. **Verification Plan** V_t is a series of tasks proposed by the MLLM to distinguish current hypotheses (e.g., verify the flag at [x, y, w, h]), and help collect new evidence, e_{t+1}. For example, a single task v^i can be defined as:

$$\begin{aligned} v^i = \{&\text{desc: "Examine sign"},\\ &\text{reason: "Sign text may provide clues"},\\ &\text{bbox: } [796, 315, 934, 357],\\ &\text{status: "Pending"}\} \end{aligned} \quad (9)$$

Global Memory M_t stores key evidence that provides high information gain for updating posterior probabilities. This allows key evidence (e.g., a Cyrillic sign) to be efficiently reused across different reasoning levels (e.g., from country to city). **Inference Context** C_t records the current reasoning level (e.g., street) and the history of hypothesis updates, which provides essential interpretability and traceability for the decision-making process.

Sequential Bayesian Updating

Built on calibrated probability distributions and evolving state memory, GeoBayes performs hierarchical geolocalization via a "Hypothesize–Verify–Update" loop grounded in Sequential Bayesian Updating. It first transforms global analysis from an MLLM into calibrated priors to initialize S_0. Then, at each step, it selects a verification task, gathers new evidence, and updates the posterior. Once the posterior surpasses a threshold, the model transitions to the next reasoning level, reusing accumulated evidence and prior context. GeoBayes progressively narrows down candidate regions and outputs the most probable location.

Hypothesis. As shown in Fig. 3, in the loop, GeoBayes first uses MLLM to perform global image analysis to generate the hypothesis set $\mathcal{H}_t$ with verification plan V_t, forming the state memory S_t.

Notably, if the image contains a definitive clue for a specific city (e.g., the Sydney Opera House), the model directly initializes at the appropriate reasoning level (e.g., city level), skipping higher levels.

Verify. Guided by S_t, GeoBayes enters the Verify stage. It selects a verification task v_i, extracts the corresponding image region, and prompts the MLLM for interpretation, yielding new evidence e_t. As in Fig. 3, for a "verify bus" task at the city level, the MLLM might return: "This is a bus with a London logo, so the area is likely located in the UK."

Update. With evidence e_t, GeoBayes quantifies its impact via likelihood estimation. First, the MLLM provides a judgment tuple (c_t, α_t), converted into a quantitative support score $W(e_t \mid l_i)$ (Eq. 6). For instance, $W(e_t \mid \cdot) = \{\text{UK} : 1.87,\ \text{US} : 0.54,\ \text{SE} : 0.56\}$. This likelihood is used to update the posterior via Bayesian updating (Eq. 7), yielding a new hypothesis set $H_t = \{\text{UK} : 0.682,\ \text{US} : 0.161,\ \text{SE} : 0.156\}$. This updated set H_t is then stored in S_t, completing one inference loop. In this case, the red bus initially increases the probability of the UK, but later evidence supports a U.S. location during Bayesian updating, effectively avoiding confirmation bias.

Enhance. Given the limitations of MLLMs in analyzing fine-grained or uncommon geographic details, GeoBayes evaluates the information gain after each hypothesis update by computing the probability change:

$$\Delta P_t = |P_t - P_{t-1}|, \quad P_t = P(L = l \mid E_{1:t}) \quad (10)$$

If $\Delta P_t < 0.05$, the current evidence is considered weak and GeoBayes will further employs a WebSearch module to gather external clues. This module includes two tools: ImageSearch, which takes a visual object O_i as input, and TextSearch, which uses a textual query. Crucially, the query for both tools is adapted based on the current reasoning level. For instance, at the country level, ImageSearch queries take the form: " O_i in which country?" At the U.S. city level, this is refined to: " O_i in which U.S. city?" If ImageSearch fails to return useful results, TextSearch is triggered using a textual description of the visual object, e.g., " Black-on-white license plate in which country?" This adaptive querying mechanism enables GeoBayes to retrieve more context-aware and level-specific information, improving evidence quality during inference.

Hierarchical Transition. GeoBayes performs multi-level reasoning, narrowing the location from country to city to street level. When the probability of the primary hypothesis exceeds the transition threshold $\tau_{transition}$, GeoBayes advances to the next level. To leverage priors from the previous level, city- or street-level hypotheses are not generated directly by the MLLM. Instead, GeoBayes calls WebSearch to retrieve external clues based on stored key evidence, generating a new city-level hypothesis set. For example, a search for "O_i(red bus) in which cities of US?" yields $H_0^{city} = \{\text{SanFrancisco} : 0.574,\ \text{LosAngeles} : 0.425\}$. Concurrently, it creates a new verification plan that prioritizes underutilized visual objects and encourages a deeper investigation of confirmed key evidence. For example, a new verification task, "check cables" focuses on transportation infrastructure. The **Verify** module returns the description: "These may be

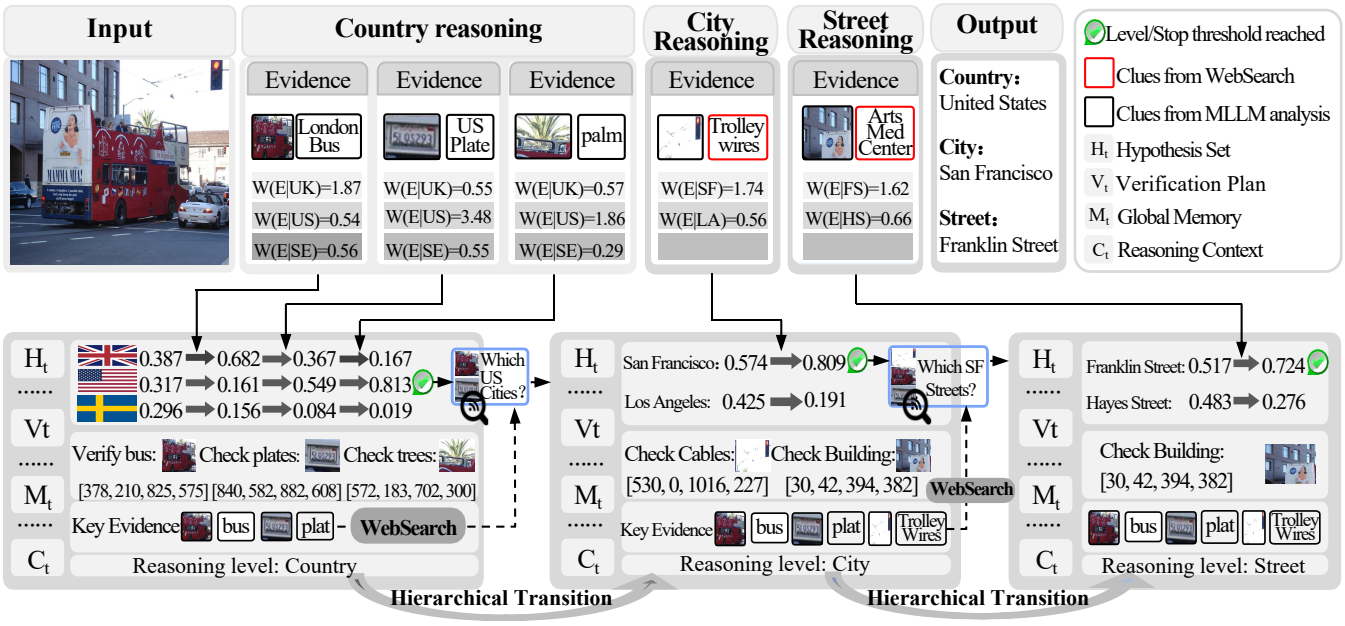

Figure 3: Example of the GeoBayes inference process. The model iteratively verifies visual clues and updates hypotheses to achieve localization from the country to the street level. Structured memory tracks hypotheses (H_t), plans (V_t), global memory (M_t), and context (C_t) across levels, enabling hierarchical transitions and coarse-to-fine localization.

dense overhead cables used to power trams", which is then quantified as evidence. This new evidence increases the posterior probability for "San Francisco," driving the sequential Bayesian updating to converge at the city level.

Replace and Backtrack. GeoBayes introduces Replace and Backtrack strategies to prevent error accumulation during hierarchical transitions. **Replace** is triggered when no hypothesis meets the confidence threshold and the current verification plan is exhausted. In this case, GeoBayes prompts the MLLM with the information that all verified candidates are invalid and asks it to regenerate a new hypothesis set and verification plan based on the previous reasoning context. **Backtracking** occurs when, at a finer-grained level, the collected evidence yields support ratings $c_t \leq 2$ for all current hypotheses. The system then halts the current process and re-evaluates country-level hypotheses in the state memory using the new evidence.

Output. GeoBayes stops inference and outputs when:

$$\text{Stop}(S_t) = \begin{cases} \max P_t(l) \geq \tau \ \wedge \ \text{Exh}(V_t), & \lambda < \text{Street}, \\ \text{Exh}(V_t), & \lambda = \text{Street}, \end{cases} \tag{11}$$

where $\max P_t(l)$ is the highest probability in the current hypothesis set, and $\text{Exh}(V_t)$ indicates that the verification plan is exhausted and no further external searches are available. λ denotes the current reasoning level (e.g., street). Inference ends either when (1) at the street level, the primary hypothesis exceeds a confidence threshold or all verification tasks are completed; or (2) at any other level, no further evidence can be gathered. Upon termination, the model outputs the location with the highest posterior (e.g., Franklin Street) via MAP inference (Eq. 4). This design enables GeoBayes to maximize precision with evidence, while producing the most promising result under limited information.

Experiments

Datasets and evaluation metrics: Following prior works (Haas et al. 2024; Jia et al. 2024), we evaluate GeoBayes on the Im2GPS3k (Hays and Efros 2008) and YFCC4K (Thomee et al. 2016) datasets. We measure localization accuracy using standard distance thresholds. Specifically, we calculate the distance between the predicted and ground-truth coordinates and report the percentage of predictions that fall within five geographical scales: Street (<1 km), City (<25 km), Region (<200 km), Country (<750 km), and Continent (<2500 km).

Implementation details: GeoBayes is model-agnostic and can operate with any capable MLLM. For simplicity, we use two open-source models: Qwen2-VL-7B and Qwen2.5-VL-7B. Following Georeasoner (Li et al. 2024) and NAVIG (Zhang et al. 2025a), we prompt the MLLM to first generate a hierarchical location name, which is then geocoded into GPS coordinates. We set the hyperparameters as follows: cutoff $\tau_p = 0.6$, transition threshold $\tau_{\text{transition}} = 0.7$ and enhancement threshold $\tau_{\text{enhance}} = 0.05$. For the WebSearch module, we employ the Google Lens API for image search and the Tavily Search API for web searches.

Comparative Results.

In this subsection, we compare GeoBayes with other methods using both quantitative and qualitative results.

1) Quantitative Results. Tab. 1 summarizes the performance of GeoBayes on Im2GPS3K and YFCC4K.

Methods	Im2GPS3K					YFCC4K				
	2500km	**750km**	**200km**	**25km**	**1km**	**2500km**	**750km**	**200km**	**25km**	**1km**
Translocator (CVPR'22)	80.1	58.9	46.7	31.1	11.8	60.4	41.1	27.0	18.6	8.4
GeoDecoder (CVPR'23)	76.1	61.0	45.9	33.5	12.8	68.7	50.0	33.9	24.4	10.3
GeoCLIP (NeurIPS'24)	83.8	69.7	50.7	34.5	14.1	74.7	55.0	32.6	19.3	9.6
MiniCPM-V (Arxiv'24)	33.2	27.8	22.4	15.9	2.3	24.3	19.9	11.3	8.0	1.3
LLaVA-Next(Arxiv'24)	61.2	43.2	25.9	16.5	2.6	50.3	28.1	14.0	7.9	1.0
Qwen2-VL (Arxiv'24)	75.0	65.0	48.9	29.9	5.3	63.0	45.9	27.3	13.8	1.9
Qwen2.5-VL (Arxiv'25)	83.8	70.4	51.1	31.0	5.1	68.8	50.7	28.1	14.1	2.1
Img2Loc-LLaVA (SIGIR'24)	51.1	40.1	29.9	23.4	8.0	39.7	30.0	19.5	14.2	**7.9**
Georeasoner (ICML'24)	80.7	62.9	43.3	28.8	**8.1**	67.3	50.6	28.5	**17.0**	3.7
NAVIG (Arxiv'25)	84.0	68.3	49.1	28.9	5.5	68.8	49.3	29.5	14.7	2.1
Geobayes (Ours)										
- Qwen2-VL-7B	80.2	69.1	49.7	31.7	6.0	69.5	49.2	28.6	14.4	3.3
- Qwen2.5-VL-7B	**85.9**	**73.7**	**53.6**	**34.7**	6.3	**75.4**	**55.8**	**30.9**	16.1	4.9

Table 1: Localization accuracy (%) on Im2GPS3K and YFCC4K. Results for classification/retrieval-based methods (top), generation-based methods (middle), and ours (bottom). Bold: best in generation-based methods.

Compared with classification/retrieval methods, GeoBayes (Qwen2.5-vl) achieves higher coarse-grained accuracy ($\geq$200 km) on Im2GPS3K. For example, it improves 200 km accuracy by 2.9% over GeoCLIP (Vivanco Cepeda, Nayak, and Shah 2023), thanks to its ability to resolve ambiguity in high-level geographic clues. However, in street-level localization, classification-based methods still perform better, as they exploit proprietary priors (e.g., store signs, house numbers), which are not included in the training of our method.

GeoBayes also outperforms generation-based methods and advanced MLLMs. Compared to Qwen2-VL, GeoBayes (Qwen2-vl) improves performance on Im2GPS3K by 5.2%, 4.1%, 0.8%, 1.8%, and 0.7% across five spatial scales, and on YFCC4K by 6.5%, 3.3%, 1.3%, 0.6%, and 1.4%. Notably, even without fine-tuning, GeoBayes (Qwen2-VL) still achieves an advantage. With Qwen2.5-VL, performance improves further. This shows GeoBayes enhances MLLMs' geospatial reasoning by dynamically integrating evidence.

2) Qualitative Results. We qualitatively compare Qwen2.5-VL-7B, GPT-4o, and GeoBayes on the same images. In Fig. 4 (top), Qwen2.5-VL-7B misidentifies the location as Washington, D.C. In contrast, GPT-4o, leveraging world knowledge, associates the image with a well-known landmark and localizes it to Barcelona. GeoBayes succeeds by using "Hypothesize-Verify-Update" loop and retrieved clues (e.g., "La Diosa"). In another challenging case where GPT-4o fails, as shown in Fig. 4 (bottom), GeoBayes utilizes an external clue (area code "01851") to localize the scene in Isle of Lewis. This indicates that a smaller MLLM is susceptible to confirmation bias, while even large MLLM can make errors in fine-grained scenarios. In contrast, GeoBayes, through iterative evidence integration, enables a smaller model to approach or match performance of a larger one without extra training.

Ablation Study.

In this subsection, we conduct ablation studies to evaluate each component of GeoBayes.

Reasoning method	750km	200km	25km	1km
COT	70.4	51.1	31.0	5.1
MCoT	71.7	52.3	26.3	4.3
MCoT + RAG	72.2	51.9	25.8	3.6
Ours	**73.7**	**53.6**	**34.7**	**6.3**

Table 2: Performance comparison of different reasoning methods on Im2GPS3k. Here, CoT uses text only; MCoT adds visual cues. All methods are based on Qwen2.5-VL.

1) Comparison of Reasoning Methods. Based on Qwen2.5-VL, we compare different reasoning methods: (1) CoT; (2) MCoT, injecting visual objects following RIV-CoT (Corbière et al. 2025); (3) MCoT+RAG, which is further augmented with image search. As shown in Tab. 2, compared to CoT, incorporating multimodal and retrieval information improves accuracy at the country level but degrades performance at finer-grained levels (city, street). This is because the mapping between visual objects and countries is relatively unambiguous, whereas at finer-grained levels such as cities, geographic ambiguity increases (e.g., similar architectural styles may appear in dozens of cities). Without fine-tuning, simply injecting visual or retrieved information into a general-purpose MLLM introduces more noise. In contrast, GeoBayes explicitly weighs each piece of evidence and updates probabilities across multiple hypotheses. This principled evidence integration suppresses noise and reinforces consistent signals, ensuring reliable localization.

2) Initial Hypothesis Reliability. We evaluate the reliability of MLLM-generated priors by computing the Top-K country-level recall for Qwen2.5-VL. As shown in Tab. 3, the Top-5 recall rates reach 81.5% on Im2GPS3k and 74.3% on YFCC4K. Notably, even on the more challenging YFCC4K dataset, the Top-5 recall remains high despite a lower Top-1 recall of 50.7%. These results show that the MLLM's embedded world knowledge already covers most ground-truth countries, providing strong priors for Bayesian updates. This supports the core design of GeoBayes, which

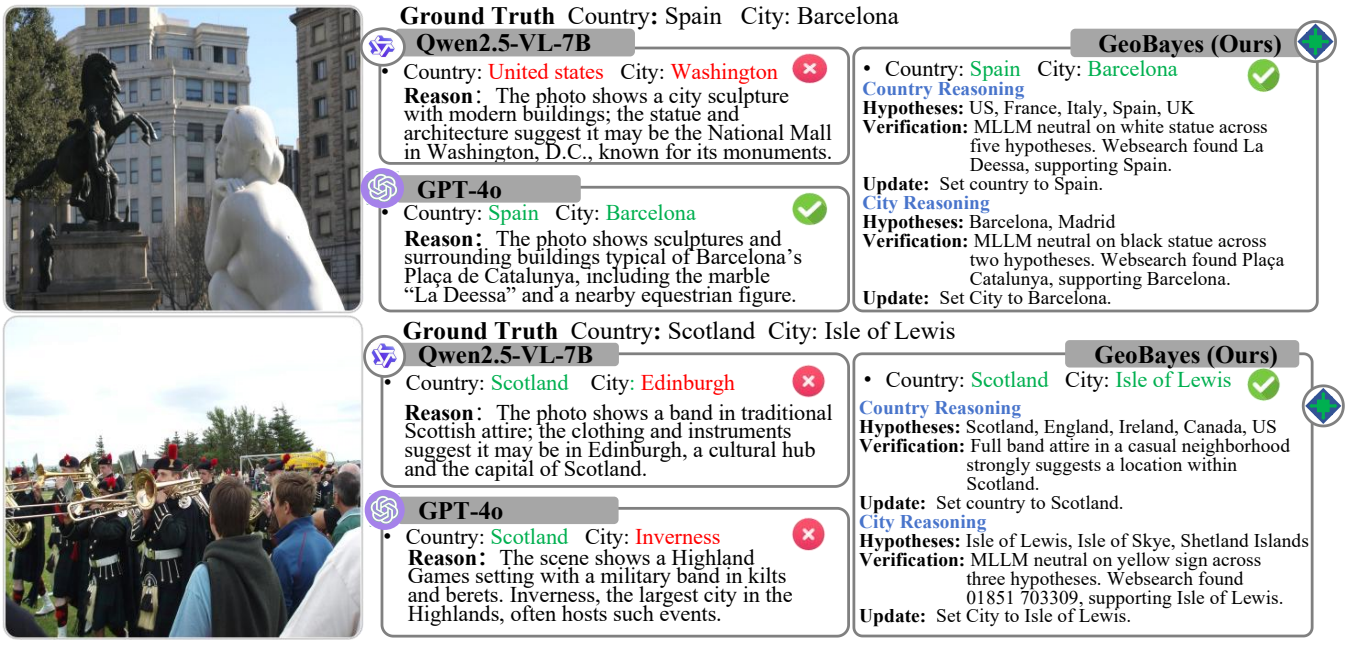

Figure 4: Qualitative comparison of geolocation across different MLLMs.

Dataset	Top-1	Top-3	Top-5
Im2GPS3K	70.4	77.6	81.5
YFCC4K	50.7	66.0	74.3

Table 3: Country-level Top-K recall(%) of initial hypotheses generated by Qwen2.5-VL on Im2GPS3K and YFCC4K.

Method	Likelihood $W(e_t \mid l)$	750km	25km	1km
w/o Calib.	$P_{LLM}(e_t \mid l)$	71.3	31.5	5.7
w/o Centering	$\exp[\alpha_t\, c_t]$	72.6	31.9	6.0
Ours	$\exp[\alpha_t \beta\,(c_t - 3)]$	**73.7**	**34.7**	**6.3**

Table 4: Effect of different likelihood formulations on Im2GPS3K. Here, w/o Calib. uses the likelihood scores returned by MLLM; w/o Center. adopts asymmetric weights.

updates probabilities across a full set of hypotheses rather than committing to a single early prediction.

3) Likelihood Estimation Strategies. As shown in Tab. 4, we compare various likelihood estimation strategies. Using raw MLLM scores (w/o Calib.) yields 31.5% accuracy within 25km. Discretizing scores into five support levels, $c_t \in \{1, \ldots, 5\}$, and applying a confidence-aware formulation $\exp[\alpha_t \cdot c_t]$ improves accuracy by 1.3% at the country level. Adding a centering term, $\ln 2(c_t - 3)$, allows positive and negative evidence to offset each other while amplifying strong positive signals. This formulation achieves the best result of 34.7% within 25km, indicating the effectiveness of the likelihood function design.

4) Ablation of State-Transition Mechanisms. GeoBayes employs a state memory for reliable long-chain reasoning. Tab. 5 evaluates the effectiveness of four key state transition strategies. First, removing hierarchical transition (w/o Hierarchy) reduces accuracy by 3.2% at 25km and 1.2% at 1km, as the lack of higher-level priors expands the candidate space and increases retrieval noise. Second, disabling retrieval enhancement (w/o Enhance) results in a drop of 3.4% at 25km, showing that external cues are crucial for resolving fine-grained ambiguity. Third, removing hypothesis updating (w/o Replace) and backtracking (w/o Backtrack) causes a modest drop, suggesting that while one-shot reasoning with multiple hypotheses captures most ground-truths, iterative updates further refine the results. These results confirm the effectiveness of the state memory mechanism.

Method	750km	200km	25km	1km
GeoBayes	**73.7**	**53.6**	**34.7**	**6.3**
w/o Hierarchy	73.2	52.0	31.5	5.1
w/o Enhance	72.8	52.5	31.3	5.3
w/o Replace	73.1	52.9	33.2	6.1
w/o Backtrack	73.7	53.1	33.7	5.8

Table 5: Performance of GeoBayes on Im2GPS3k under different state transition control strategies.

Conclusion

This paper proposes GeoBayes, which formulates image geo-localization as a MAP estimation problem solved via sequential Bayesian updating. GeoBayes uses a "Hypothesize-Verify-Update" loop to progressively fuse geographic evidence across spatial levels. Built on Qwen2.5-VL, GeoBayes improves average localization accuracy by 3.2% on the public datasets, outperforming multiple generation-based baselines without any additional training.

Acknowledgments

This paper is supported by the Science and Technology Project of Hainan Provincial Department of Transportation (Grant No.HNJTT-KXC-2024-3-22-02), the National Natural Science Foundation of China (Grant No. 62272018, 62206184), and Zhongguancun Laboratory.

References

Cheng, A.-C.; Yin, H.; Fu, Y.; Guo, Q.; Yang, R.; Kautz, J.; Wang, X.; and Liu, S. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37: 135062–135093.

Clark, B.; Kerrigan, A.; Kulkarni, P. P.; Cepeda, V. V.; and Shah, M. 2023. 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*, 23182–23190.

Corbière, C.; Roburin, S.; Montariol, S.; Bosselut, A.; and Alahi, A. 2025. Retrieval-based interleaved visual chain-of-thought in real-world driving scenarios. *arXiv preprint arXiv:2501.04671*.

Gao, J.; Li, Y.; Cao, Z.; and Li, W. 2025. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19520–19529.

Haas, L.; Skreta, M.; Alberti, S.; and Finn, C. 2024. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12893–12902.

Han, J.; Buntine, W.; and Shareghi, E. 2024. Towards uncertainty-aware language agent. *arXiv preprint arXiv:2401.14016*.

Han, X.; Zhu, C.; Zhao, X.; and Zhu, H. 2024. Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. *arXiv preprint arXiv:2408.11312*.

Hays, J.; and Efros, A. A. 2008. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, 1–8. IEEE.

Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 22105–22113.

Jia, P.; Liu, Y.; Li, X.; Zhao, X.; Wang, Y.; Du, Y.; Han, X.; Wei, X.; Wang, S.; and Yin, D. 2024. G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. *Advances in Neural Information Processing Systems*, 37: 53198–53221.

Jiang, X.; Fang, Y.; Qiu, R.; Zhang, H.; Xu, Y.; Chen, H.; Zhang, W.; Zhang, R.; Fang, Y.; Chu, X.; et al. 2024. Tc-rag: turing-complete rag's case study on medical llm systems. *arXiv preprint arXiv:2408.09199*.

Li, C.; Wu, W.; Zhang, H.; Xia, Y.; Mao, S.; Dong, L.; Vulić, I.; and Wei, F. 2025. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*.

Li, L.; Ye, Y.; Jiang, B.; and Zeng, W. 2024. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Forty-first International Conference on Machine Learning*.

Liu, L.; and Li, H. 2019. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5624–5633.

Mandal, S. 2024. A privacy preserving federated learning (PPFL) based cognitive digital twin (CDT) framework for smart cities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23399–23400.

Manvi, R.; Khanna, S.; Mai, G.; Burke, M.; Lobell, D.; and Ermon, S. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*.

Mitra, C.; Huang, B.; Darrell, T.; and Herzig, R. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14420–14431.

Muller-Budack, E.; Pustu-Iren, K.; and Ewerth, R. 2018. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)*, 563–579.

Pramanick, S.; Nowara, E. M.; Gleason, J.; Castillo, C. D.; and Chellappa, R. 2022. Where in the world is this image? transformer-based geo-localization in the wild. In *European Conference on Computer Vision*, 196–215. Springer.

Seo, P. H.; Weyand, T.; Sim, J.; and Han, B. 2018. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 536–551.

Shi, Y.; Yu, X.; Campbell, D.; and Li, H. 2020. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4064–4072.

Theiner, J.; Muller-Budack, E.; and Ewerth, R. 2022. Interpretable semantic photo geolocation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 750–760.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

Vivanco Cepeda, V.; Nayak, G. K.; and Shah, M. 2023. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36: 8690–8701.

Vo, N.; Jacobs, N.; and Hays, J. 2017. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*, 2621–2630.

Wang, J.; Zheng, Z.; Chen, Z.; Ma, A.; and Zhong, Y. 2024. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 5481–5489.

Wang, Y.; Wang, S.; Cheng, Q.; Fei, Z.; Ding, L.; Guo, Q.; Tao, D.; and Qiu, X. 2025. Visuothink: Empowering lvlm reasoning with multimodal tree search. *arXiv preprint arXiv:2504.09130*.

Weyand, T.; Kostrikov, I.; and Philbin, J. 2016. Planet-photo geolocation with convolutional neural networks. In *European conference on computer vision*, 37–55. Springer.

Wu, W.; Mao, S.; Zhang, Y.; Xia, Y.; Dong, L.; Cui, L.; and Wei, F. 2024. Mind's eye of LLMs: visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems*, 37: 90277–90317.

Xu, S.; Zhang, C.; Fan, L.; Meng, G.; Xiang, S.; and Ye, J. 2024. Addressclip: Empowering vision-language models for city-wide image address localization. In *European Conference on Computer Vision*, 76–92. Springer.

Yang, H.; Lu, X.; and Zhu, Y. 2021. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34: 29009–29020.

Zemene, E.; Tesfaye, Y. T.; Idrees, H.; Prati, A.; Pelillo, M.; and Shah, M. 2018. Large-scale image geo-localization using dominant sets. *IEEE transactions on pattern analysis and machine intelligence*, 41(1): 148–161.

Zhang, Z.; Li, R.; Kabir, T.; and Boyd-Graber, J. 2025a. Navig: Natural language-guided analysis with vision language models for image geo-localization. *arXiv preprint arXiv:2502.14638*.

Zhang, Z.; Zhou, W.; Zhao, J.; and Li, H. 2025b. Robust Multimodal Large Language Models Against Modality Conflict. *arXiv preprint arXiv:2507.07151*.

Zhou, Z.; Zhang, J.; Guan, Z.; Hu, M.; Lao, N.; Mu, L.; Li, S.; and Mai, G. 2024. 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*, 2749–2754.

Zhu, S.; Shah, M.; and Chen, C. 2022. 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*, 1162–1171.

Zhu, S.; Yang, L.; Chen, C.; Shah, M.; Shen, X.; and Wang, H. 2023. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19370–19380.

Zhu, S.; Yang, T.; and Chen, C. 2021. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3640–3649.