Improving VLM Geo-Localization with Retrieval and Reasoning

Bofeng Cao[†], Nick Boddy[*], Prajjwal Gandharv[*], Riyad Hassen[*]
Author names are listed in alphabetical order.
Department of Computer Sciences[*], Electrical and Computer Engineering[†], University of Wisconsin–Madison

Introduction

Image geo-localization is the task of deciding geographic location of an image using only its visual cues. Immediate applications span mapping and navigation, augmented reality, media management (e.g., clustering and tagging photos by location), crisis response, and digital forensics.

At the same time, the ability to localize everyday photos raises safety, privacy, and misinformation concerns. Accurate models can infer where people live, work, or travel, from seemingly benign social-media posts, which enables unsolicited tracking or even targeted harm. Fake or out-of-context images can also be used to misrepresent where events occurred, making fabricated evidence harder to spot at scale.

In this project, we study VLM-based geo-localization in two directions: (i) detecting geographically-inconsistent or fake images, and (ii) augmenting existing geo-localization methods with reasoning (CoT) and retrieval (RAG).

Figure 1. GeoGuessr: geo-localization framed as an online game, where players compete for score measured by distance from target. The 2013 game found a surge in popularity in 2021 which lended to a wide development of interest and "expertise" in geo-localization.

Background & Related Work

With the rapid development of vision-language models (VLMs), a growing body of research has begun to explore their capabilities for geo-localization, achieving impressive progress in recent years [1, 2, 3, 4, 5]. These methods leverage the strong semantic understanding and cross-modal alignment of VLMs to infer geographic clues from street-view imagery.

Among them, **GeoRanker**[3] stands out by introducing a distance-aware ranking framework for fine-grained geo-localization, achieving state-of-the-art performance in predicting precise geographic coordinates.

On the other hand, **GeoReasoner**[4] and subsequent work by Liu et al.[1] emphasize the reasoning abilities of VLMs in geo-localization. These approaches not only predict the final geographic position but also elicit explicit reasoning chains that justify the prediction.

Motivation & Problem

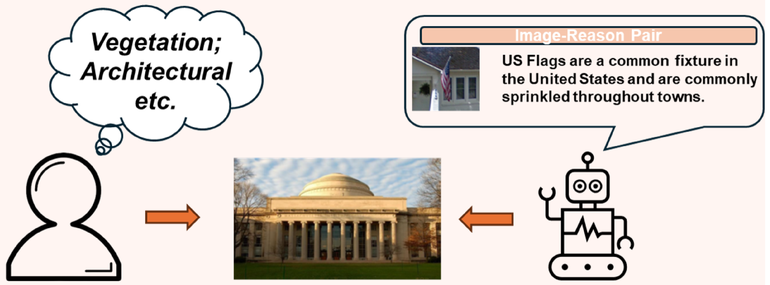

Figure 2. Expert vs. VLM

An Expert GeoGuessr[6] player typically identifies location cues by integrating diverse visual evidence—such as vegetation types, climatic signatures, terrain and landform characteristics, architectural styles, and their spatial relationships—to narrow down the geographic search space.

In contrast, it is uncertain whether current VLMs follow comparable reasoning processes or whether their apparent success primarily stems from memorization and output regularization.

Due in part to this motivation, we create the *GeoConflict* dataset for a new task that we believe may improve generalization to geo-localization.

- **GeoConflict.** We use a VLM-driven image editing model to inject *conflicting geographic attributes* into real street-view images. An example might be replacing a coniferous tree in a Wisconsin scene with a palm tree.
- **Geographic Self-Consistency.** Decide/reason whether a given image is self-consistent with its geography, or if it is fake.
- **GeoConflict fine-tuning.** We replace 10 percent of the original GeoReasoner training data with GeoConflict samples and fine-tune the model to improve its ability to resolve geographic inconsistencies.

Example Results

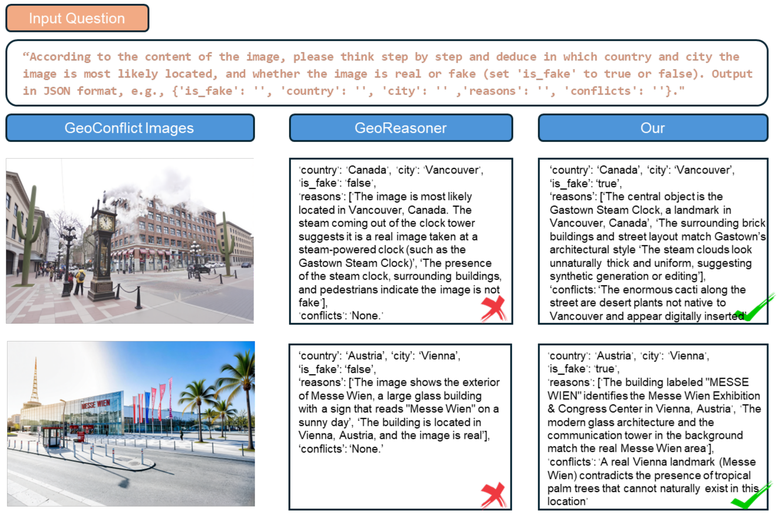

Figure 3. Results Comparison between GeoReasoner and Our's on a GeoConflict task.

Dataset

NaviClues - includes panoramic image, (lat,long) pairs, and reasoning.

Guidebooks - contains an image-text pair from the *Plonk It* Guide to GeoGuessr website. We used it for training a CLIP encoder for the RAG pipeline.

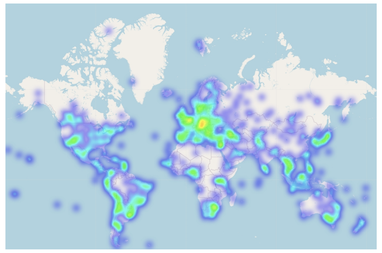

Figure 4. Location distribution of NaviClues, covering a wide range of countries around the world.

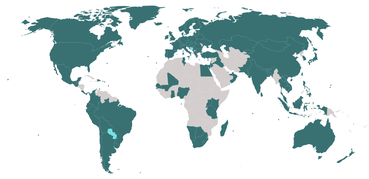

Figure 5. Coverage of the Plonk It community guide as of November 2025. Cyan color indicates work in progress.

Methodology

We follow a similar strategy used in the NAVIG, ETHAN, and GeoReasoner paper. There are three main components:

1. **Reasoner** - Given Image I, a VLM generates reasoning R, which includes a textual description of the image.

2. **Retrieval Stage** - We encode images from the *Plonk It* dataset with the CLIP model for the purpose of retrieving similar image locations from the database. FAISS (Facebook AI Similarity Search) as a vector database for the RAG pipeline with L_2 distance for retrieving from the vector database $d = \arg\min \|x - x_i\|_2$.

3. **Prediction Stage** - In conjunction with the retrieved information from stages 1 and 2, the original image is passed to a VLM to predict the final location. $\hat{Y} := \text{VLM}_P(I, \text{concat}(R, K))$.

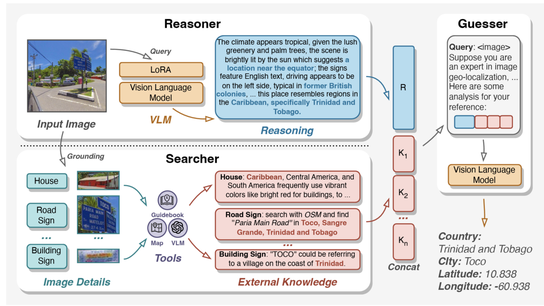

Figure 6. NAVIG Architecture

GeoConflict Generation

To construct the GeoConflict dataset, we begin by providing a real-world street-view (GSV) image to a VLM. The VLM is prompted to identify and describe the geographic cues present in the scene. These cues span multiple categories, including architectural style, landform features, vegetation types, transportation infrastructure, and cultural or commercial indicators.

Next, we randomly select one geographic cue from a chosen category and replace it with a conflicting cue. This replacement cue is determined by our custom RAG system, which integrates (1) a classifier that categorizes geographic cues and (2) a fine-tuned CLIP model that retrieves plausible yet contradictory cues.

Finally, the conflicting cue is converted into a text prompt and, together with the original street-view image, is sent to a VLM-based image editing model. This model generates the modified image, resulting in a GeoConflict sample that intentionally embeds mismatched or geographically inconsistent elements.

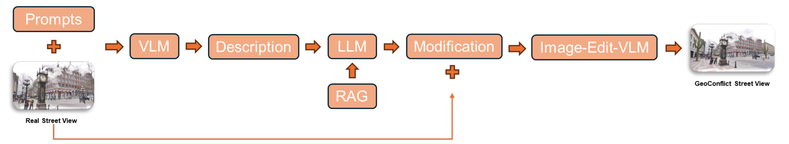

Figure 7. Pipeline of GeoConflict Data Generation

Experiment Details

We conduct experiments on Geo-Localization and Geographic Self-Consistency following the original GeoReasoner setup, and use Qwen-VL-Chat as our baseline VLM. To construct the GeoConflict dataset, we use GPT-4o for image description and modification-prompt generation, and qwen-image-edit as the image editing model. The model selections are summarized in Table 1.

Component / Task	Model Used	Purpose
Baseline VLM	Qwen-VL-Chat	Geo-Localization and Geographic Self-Consistency
Fine-tuning backbone	Qwen-VL-Chat	Fine-tuned with 10 percent GeoConflict samples
Image description / Modification	GPT-4o	Generate modification prompts for GeoConflict
Image editing model	qwen-image-edit	Inject conflicting geographic attributes

Table 1. Model configurations used in our experiments.

We conducted all training and testing on Nvidia Ada 6000 GPUs (48GB) using CUDA 12.1, PyTorch 2.3.1, and Transformers 4.33.0. The detailed training configurations for our GeoReasoner variants are summarized in Table 2.

Model Variant	Training Speed	Inference Latency	#Params	LoRA Params	FLOPs
LoRA1 (reason)	0.3 samples/s	2.20 s	9.6B	112.19M	71.9B
LoRA2 (location)	0.4 samples/s	1.20 s	9.6B	112.19M	71.9B

Table 2. Training details of the proposed GeoReasoner variants.

Limitations & Future Work

Our current approach has several limitations.

- Eliciting reasoning through free-form explanations makes it difficult to quantitatively assess the model's true geographic reasoning ability.
- Our current evaluation focuses mainly on city-level localization and reasoning about natural vegetation, leaving broader geographic contexts underexplored.

To address these limitations, we plan to incorporate Process Reward Models[7, 8] (PRMs) to provide a more principled and quantitative evaluation framework. Rather than asking the model to directly output a single "reason", we will generate multiple Chain-of-Thought (CoT) traces and score them step-by-step using a PRM. Formally, given a CoT $S = (s_1, s_2, \ldots, s_k)$ consisting of k reasoning steps, a PRM maps the entire sequence to a k-dimensional reward vector $\text{PRM}(S) \in [0, 1]^k$, where the i-th component $\text{PRM}(S)_i$ represents the predicted correctness of step s_i. This allows us to evaluate geographic reasoning in a more fine-grained and scalable manner.

References

[1] Y. Liu, G. Deng, J. Ding, Y. Li, T. Zhang, W. Sun, Y. Zheng, and J. Ge, "Mission: Impossible - image-based geolocation with large vision-language models," *Proceedings on Privacy Enhancing Technologies*, vol. 2025, no. 4, pp. 410–428, 2025.

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

[3] P. Jia, S. Park, S. Gao, X. Zhao, and Y. Li, "Georanker: Distance-aware ranking for worldwide image geolocalization," *arXiv preprint arXiv:2505.13731*, 2025.

[4] L. Li, Y. Ye, and W. Zeng, "Georeasoner: Geo-localization with reasoning in street views using a large vision-language model," in *International Conference on Machine Learning (ICML)*, 2024.

[5] Z. Zhou, J. Zhang, Z. Guan, M. Hu, N. Lao, L. Mu, S. Li, and G. 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*, pp. 2749–2754, 2024.

[6] "Geoguessr." `https://www.geoguessr.com/`. Accessed: 2025-02-11.

[7] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," 2023.

[8] T. Zeng, S. Zhang, S. Wu, C. Classen, D. Chae, E. Ewer, M. Lee, H. Kim, W. Kang, J. Kunde, Y. Fan, J. Kim, H. I. Koo, K. Ramchandran, D. Papailiopoulos, and K. Lee, "Versaprm: Multi-domain process reward model via synthetic reasoning data," in *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.