# GeoAgent: Precise Worldwide Multimedia Geolocation with Large Multimodal Models

**Jett Chen**
Shanghai Starriver Bilingual School
jettchen.t@outlook.com

## Abstract

Precise worldwide geolocalization is a challenging task due to the variety of images taken from different geographical and cultural landscapes. Existing work on world-wide geolocalization uses machine-learning approaches to estimate the location of images by constructing mappings between imagery data and their corresponding coordinates or geographical regions. As such, existing models cannot leverage online or contextual information related to their task, and are unable to effectively make precise street-level(1km) predictions. We present GeoAgent, a Large Multi-modal Model(LMM) powered agent that completes precise geolocalization tasks by leveraging online information such as Open Streetmap, satellite imagery, Google Street View, as well as other machine learning models for image retrieval and geo-estimation. We evaluate our model on the image geolocalization benchmark IM2GPS3k and find that it outperforms the previous SOTA in street-level geolocalization by an accuracy of $20.7\%$. Overall, we demonstrate that multi-modal vision agents' potential to be applied to real-world, complex tasks that require reasoning.

## 1 Introduction

Geolocalization, or geolocation, refers to the task of determining the geographical location associated with a piece of media. Though popularized by the game Geo-Guessr, geolocation has a variety of real-world use cases such as crisis response and verification of online content. Geolocalization involves the usage of all forms of content, such as news articles, tweets, and image metadata, to identify and corroborate the location associated with such content. Geolocation is considered one of the most time-consuming and complex steps in the open source intelligence and digital verification process (Fred et al., 2018). Due to the variety of geographical and cultural regions on Earth, precise world-wide geolocalization remains a challenging task.

Existing work on geolocalization primarily focuses on image-based retrieval methods, which match query images with a database of geographically annotated images, demonstrating high retrieval accuracy(Lu et al., 2024; Ali-bey et al., 2023; Berton et al., 2023; Deuser et al., 2023; Zhu et al., 2022, 2023; Shi et al., 2019). However, the requirement of a comprehensive global database makes retrieval-based methods impractical for world-wide geolocalization. Instead, world-wide geolocalization relies on classification or transformer models to map imagery data to geographical regions(Hays & Efros, 2008; Weyand et al., 2016; Seo et al., 2018; Clark et al., 2023; Haas et al., 2023; Clark et al., 2023; Pramanick et al., 2022; Cepeda et al., 2023). These state-of-the-art models recognize features like vegetation and road markings, similar to professional GeoGuessr players. However, unlike constrained GeoGuessr gameplay, digital investigators use external tools such as Google Earth and satellite imagery for precise geolocations. Current methods lack the ability to incorporate these external sources, limiting their precision at the street level.

Recent advancements in large-multimodal models enables models to dynamically retrieve and process online information, presenting new opportunities for more precise geolocation(Yang et al., 2023b;

Preprint. Under review.

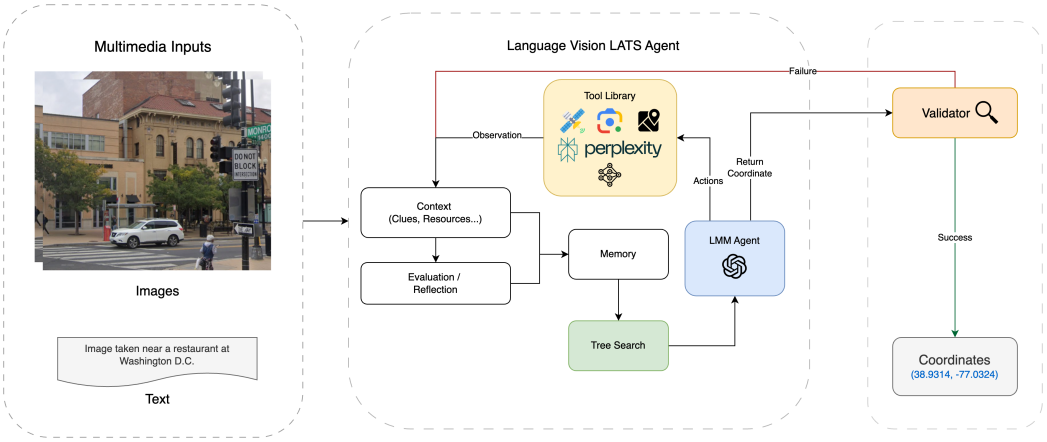


Figure 1: GeoAgent is a multi-modal agent based on Language Agent Tree Search that specializes in conducting precise world-wide geolocalization. GeoAgent has access to tools such as Google Maps, satellite imagery, Perplexity, and more

Wu et al., 2023; Niu et al., 2024; Yang et al., 2023a). To address the limitations of existing methods, we present GeoAgent, a multi-modal agent for precise world-wide geolocalization. GeoAgent uses Language Agent Tree Search(Zhou et al., 2023) to orchestrate and manage the process of a geolocation investigation, and performs actions through tools that can take multi-modal inputs and outputs. GeoAgent has access to information retrieval tools such as Google Streetview, Satellite imagery, Google Lens, Perplexity, as well as machine learning models for region estimation and retrieval-based geolocation such as GeoCLIP(Cepeda et al., 2023), EigenPlaces(Berton et al., 2023), and Sample4Geo(Deuser et al., 2023). We evaluated GeoAgent on existing image-based geolocation benchmarks and found that it outperforms the state-of-the-art model in street-level (1km) geolocalization.

We summarize our major contributions are as follows:

- We propose GeoAgent, the first language vision agent-based system that solve the problem of world-wide geolocalization, and adapted agent reasoning framework Language Agent Tree Search(Zhou et al., 2023) to the problem of world-wide geolocalization.
- We evaluated GeoAgent on the existing image-based geolocalization benchmark IM2GPS 3k, demonstrating a significant improvement in geolocalization accuracy over current state-of-the-art models.

## 2 Proposed Approach

GeoAgent is based on Language Agent Tree Search(Zhou et al., 2023), which adapts the Monte Carlo Tree Search algorithm in reinforcement learning as a framework for language model-based agents, with a focus on being general, deliberate, adaptable, flexible and modular. In LATS, the root node of the tree search represents user input. Each state transition represents an action sampled from the action space by the LM(Language Model). Thus, each node or state on the tree represents a unique trajectory or prompt that the LM can use to generate new actions and child states. GeoAgent assumes an underlying LM that is multi-modal, and all tools in GeoAgent may have multi-modal outputs in the form of text and image. LATS consists of 6 operations performed in succession: *selection, expansion, evaluation, simulation, backpropagation, and reflection.* The following will describe each of the steps in context of GeoAgent.

**Selection.** During the selection stage, the tree is traversed starting from the root node, each time selecting the child node with the highest UCT (Upper confidence bounds applied to Trees)value (Kocsis & Szepesvári, 2006) until a leaf node is selected.

**Expansion.** After selection, we prompt the LM to sample up to $N$ actions with the context of all previous actions taken in the current branch and the global investigation context. The generated actions will then be executed in parallel. The tool library is defined as follows:

- Reasoning and Control Tools:
  - `Return_coordinates(lat, lon)`: Sends coordinates to the verifier module upon location conclusion.
  - `Decide(decision)`: Records a persistent decision in its branch.
  - `Add_clue(clue)`: Adds a clue to the global investigation context.
  - `Save_coords(coords)`: Stores coordinates (in JSON) locally for further analysis by different tools.
- External Information Retrieval Tools:
  - `Google_maps_search(location)`: Searches Google Maps for a location, returning coordinates and metadata.
  - `Google_lens(image_id)`: Uses Google Lens to find visual matches for the image.
  - `Google_image_search(query)`: Searches Google Images for the provided query.
  - `Perplexity_ask(query)`: Queries `perplexity.ai` for real-time information.
  - `Get_streetviews(coords_path)`: Samples up to 120 Google Street View panoramas near the coordinates, returning a database of panoramas and a render of up to 15 street views.
  - `Plot_satellite(coords_path)`: Samples and renders satellite imagery from the provided coordinates.
- ML Inference Tools:
  - `Geoclip_predict(image_id)`: Runs GeoCLIP inference on the image, returning the top 5 locations and a map with pins (Cepeda et al., 2023).
  - `Streetview_locate(image_id, db_path)`: Runs EigenPlaces model on the image using a Street View database (Berton et al., 2023).
  - `Satellite_locate(image_id, db_path)`: Runs Sample4Geo model on the image using a satellite imagery database (Deuser et al., 2023).

**Evaluation.** In this step, the LM will evaluate the trajectory of each generated action, and output a value indicating the degree of confidence that the current trajectory will lead to a successful geolocation. Denoting the currently selected node as $p$ and $s_1, s_2, ..., s_n$ as the $n \leq N$ child nodes added to the tree during expansion expansion. The LM is given a concatenation of the context of $p$ and the $n$ action-observation pairs of the child nodes, and then prompted to reason and compare the trajectory of each of the child nodes, and finally output $n$ integers each representing a value of a child node.

**Simulation.** During simulation, the currently selected node is expanded and a new child node is selected until a terminal state is reached. A terminal state is reached when `Return_coordinates` is called with a valid set of coordinates. While LATS assumes that reaching the terminal state allows the agent to objectively determine the correctness of a trajectory, such assumption could not be made for the context of geolocation. Thus, we use the LM as a judge for whether the conclusion of the investigation is correct. Since valid coordinates must be provided in order to reach a terminal state, satellite and street view imagery would be optimistically provided to the judging LM to assist in corroboration. The LM would give a score on the scale of 1 to 10, in which a 10 would indicate a successful geolocation.

**Backpropagation.** Based on the result of the simulation, backpropagation would update the values of the tree. These values would then be used in the UCT formula to guide the selection of the next node.

**Reflection.** Upon reaching an unsuccessful terminal node, the LM is prompted with the corresponding trajectory to generate (1) a summary of the attempted geolocalization process, (2) an analysis of the reasoning trajectory, and (3) advice to aid future geolocalization attempts.

While generally following the construction of Language Agent Tree Search, GeoAgent differs from LATS in the process of expansion, the evaluation of terminal nodes, and the persistence of reflections to better adapt to the task of world-wide geolocation.

Table 1: We compare the performance of GeoAgent with state-of-the-art models on the IM2GPS3k dataset (Hays & Efros, 2008)

| Method | Street 1 km | City 25 km | Region 200 km | Country 750 km | Continent 2500 km |
|---|---|---|---|---|---|
| [L]kNN, $\sigma = 4$ (Vo et al., 2017) | 7.2 | 19.4 | 26.9 | 38.9 | 55.9 |
| PlaNet (Weyand et al., 2016) | 8.5 | 24.8 | 34.3 | 48.4 | 64.6 |
| CPlaNet (Seo et al., 2018) | 10.2 | 26.5 | 34.6 | 48.6 | 64.6 |
| ISNs (Müller-Budack et al., 2018) | 10.5 | 28.0 | 36.6 | 49.7 | 66.0 |
| Translocator (Pramanick et al., 2022) | 11.8 | 31.1 | 46.7 | 58.9 | 80.1 |
| GeoDecoder (Clark et al., 2023) | 12.8 | 33.5 | 45.9 | 61.0 | 76.1 |
| GeoCLIP (Cepeda et al., 2023) | 14.11 | 34.47 | 50.65 | 69.67 | 83.82 |
| PIGEOTTO (Haas et al., 2023) | 11.3 | 36.7 | 53.8 | 72.4 | **85.3** |
| **Ours (Sampled n=50)** | **32.0** | **57.0** | **72.0** | **80.0** | 80.0 |

## 3 Evaluation

We compare the performance of GeoAgent against existing models on the task of worldwide geolocalization. For our evaluation, we use GPT-4v(OpenAI, 2023) as our base vision-language model, and we use a branch count of 5 and rollout threshold of 8 for the LATS process. Due to the significant costs of running inference on large multi-modal models, we evaluated the image geolocalization accuracy of GeoAgent by randomly sampling 50 images from the IM2GPS3k dataset. 1 compares the efficiency of GeoAgent against existing worldwide image geolocalization models. In particular, our approach significantly improves on the prior state-of-the-art performance on street, city, region, and country level geolocalization with accuracy improvements of 17.9%, 21.3%, 18.2%, 7.5%

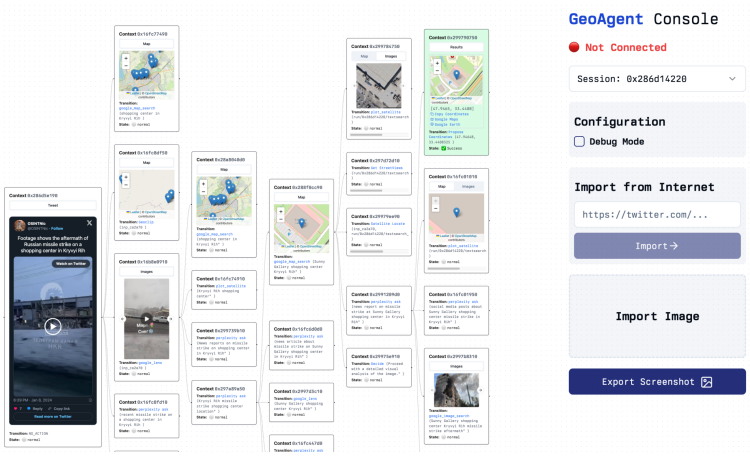


Figure 2: Visualization of GeoAgent's investigation on a tweet using GeoAgent Client. GeoAgent uses Google Lens on a video frame to identify the location; and uses perplexity and satellite imagery to corroborate the location before reaching a final conclusion. Original tweet: `http://archive.today/NrTBv`, Full screenshot of reasoning trace: `https://sh.jettchen.me/ga-trace-1`

## 4 Conclusion

In this work, we introduce GeoAgent, a multi-modal agent system for world-wide geolocalization based on Language Agent Tree Search. Through a sampled evaluation on image-based geolocalization dataset, we show that GeoAgent exhibits significant improvements over existing state-of-the-art models in terms of precise geolocalization accuracy. GeoAgent highlights the potential of multimodal agents and assistants in conducting geolocations in a real-world investigative setting.

The source code for this project is available at https://github.com/JettChenT/GeoAgent. The repo contains the core agent system under `src`, and a graph-based client under `client`. Full demo of an investigation and screenshots are available at `https://sh.jettchen.me/geoagent-demo`.

## References


Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Mixvpr: Feature mixing for visual place recognition. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2997–3006, 2023. URL https://api.semanticscholar.org/CorpusID:256646815.

Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlos German Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11046–11056, 2023. URL https://api.semanticscholar.org/CorpusID:261049627.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *arXiv preprint arXiv:2309.16020*, 2023.

Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes, 2023.

Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16847–16856, 2023.

Aahsberg Fred, Dolecek Jenna, Groenendijk Mark, and Olivia Iannelli. Introductory guide to open source intelligence and digital verification. 2018. URL https://www1.essex.ac.uk/hrc/documents/Introductory_Guide_to_Open_Source_Inteligence_and_Digitial%20Verification.pdf.

Lukas Haas, Silas Alberti, and Michal Skreta. Pigeon: Predicting image geolocations. *arXiv preprint arXiv:2307.05845*, 2023.

James Hays and Alexei A. Efros. Im2gps: estimating geographic information from a single image. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008. URL https://api.semanticscholar.org/CorpusID:2061602.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou (eds.), *Machine Learning: ECML 2006*, pp. 282–293, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-46056-5.

Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, volume 11216 of *Lecture Notes in Computer Science*, pp. 575–592. Springer, 2018. doi: 10.1007/978-3-030-01258-8_35. URL https://doi.org/10.1007/978-3-030-01258-8_35.

Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. Screenagent: A vision language model-driven computer control agent. *arXiv preprint arXiv:2402.07945*, 2024.

OpenAI. Gpt-4 vision systems card. Technical report, 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pp. 196–215. Springer, 2022.

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

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

Chenfei Wu, Sheng-Kai Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *ArXiv*, abs/2303.04671, 2023. URL https://api.semanticscholar.org/CorpusID:257404891.

Zhao Yang, Jiaxuan Liu, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. *arXiv preprint arXiv:2312.13771*, 2023a.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *ArXiv*, abs/2303.11381, 2023b. URL https://api.semanticscholar.org/CorpusID:257637012.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. *ArXiv*, abs/2310.04406, 2023. URL https://api.semanticscholar.org/CorpusID:263829963.

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

Yingying Zhu, Hongji Yang, Yuxin Lu, and Qiang Huang. Simple, effective and general: A new backbone for cross-view image geo-localization. *arXiv preprint arXiv:2302.01572*, 2023.