Swarm Intelligence in Geo-Localization: A Multi-Agent Large Vision-Language Model Collaborative Framework

Xiao Han
hahahenha@gmail.com
City University of Hong Kong;
Zhejiang University of Technology
Hong Kong/Hangzhou, China

Chen Zhu*
zc3930155@gmail.com
University of Science and Technology of China
Hefei, China

Hengshu Zhu*
zhuhengshu@gmail.com
Computer Network Information Center, Chinese Academy
of Sciences; University of Chinese Academy of Sciences
Beijing, China

Xiangyu Zhao*
xianzhao@cityu.edu.hk
City University of Hong Kong
Hong Kong, China

Abstract

Visual geo-localization demands in-depth knowledge and advanced reasoning skills to associate images with precise real-world geographic locations. Existing image database retrieval methods are limited by the impracticality of storing sufficient visual records of global landmarks. Recently, Large Vision-Language Models (LVLMs) have demonstrated the capability of geo-localization through Visual Question Answering (VQA), enabling a solution that does not require external geo-tagged image records. However, the performance of a single LVLM is still limited by its intrinsic knowledge and reasoning capabilities. To address these challenges, we introduce smileGeo, a novel visual geo-localization framework that leverages multiple Internet-enabled LVLM agents operating within an agent-based architecture. By facilitating inter-agent communication, smileGeo integrates the inherent knowledge of these agents with additional retrieved information, enhancing the ability to effectively localize images. Furthermore, our framework incorporates a dynamic learning strategy that optimizes agent communication, reducing redundant interactions and enhancing overall system efficiency. To validate the effectiveness of the proposed framework, we conducted experiments on three different datasets, and the results show that our approach significantly outperforms current state-of-the-art methods. The source code is available at https://github.com/Applied-Machine-Learning-Lab/smileGeo.

CCS Concepts

• **Information systems** → **Multimedia information systems**; • **Computing methodologies** → **Intelligent agents**.

* Chen Zhu, Hengshu Zhu, and Xiangyu Zhao are the corresponding authors.

Keywords

Retrieval-augmented LLM, Collaboration Social Network, Graph Neural Networks, Large Vision-Language Model, Geo-Localization

1 Introduction

Visual geo-localization, referred to the task of estimating geographical identification for a given image, is vital in various fields, such as analyzing historical human mobility patterns [13, 16, 28, 48] and providing location-aware, city-level attraction recommendations [5, 11, 15, 23]. In general, accurately geo-localizing images without relying on localization-based metadata (e.g., GPS tags) is a complex task that demands extensive geospatial knowledge and advanced reasoning capabilities. Traditional methods [4, 6, 8, 49] typically formulate it as an image retrieval problem where to geo-localize the given image by retrieving similar images with known geographical locations. Thus, their effectiveness is limited by the scope and quality of the geo-tagged image records.

A straightforward approach to mitigate the limitation of these databases is to deploy an agent-based framework, leveraging swarm intelligence across multiple retrieval systems [3, 35]. However, integrating diverse and independent traditional retrieval systems as agents within a unified framework presents significant challenges without human intervention. For instance, as illustrated in Figure 1, two distinct retrieval systems may provide different answers for the same input image, making it difficult to decide on the correct response without third-party mediation. This challenge highlights the inherent difficulty of coordinating independent systems within a swarm intelligence framework.

In parallel with agent-based approaches, recent advancements in Large Vision-Language Models (LVLMs) have opened up new possibilities for multi-modal tasks, such as Visual Question Answering (VQA) [9, 41]. LVLMs offer an innovative solution to visual geo-localization without relying on external geo-tagged image records.
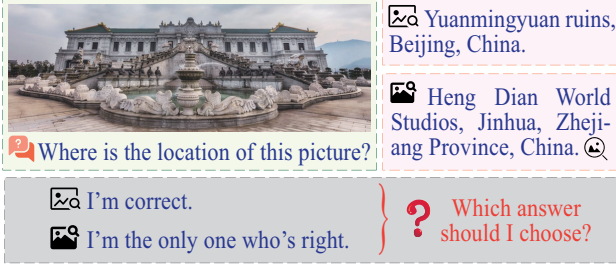
Figure 1: A toy example in the traditional agent-based retrieval framework.

Additionally, enabling LVLMs to autonomously query network search interfaces for information retrieval can further enhance their capabilities [45]. However, while individual LVLMs possess strong reasoning abilities, they still struggle with fine-grained recognition across diverse and complex scenes [20, 26, 46].

To address the limitations of both traditional retrieval systems and individual LVLMs, we propose a novel multi-LVLM agent framework, named **s**war**m** **i**ntel**lige**nce **Geo**-localization (**smileGeo**). It leverages the swarm intelligence of multiple LVLMs, each equipped with network retrieval capabilities, to collaboratively and efficiently geo-localize images. Specifically, for a given image, the framework initially selects K appropriate LVLM agents to serve as answer agents responsible for conducting the initial location analysis. Then multiple review agents are assigned to each answer agent through an adaptive social network, simulating the collaborative relationships between agents, to refine its analysis within the visual geo-localization task. Following this, the framework facilitates open discussions among all answer agents to reach a consensus. However, as the number of agents grows, managing these discussions can become increasingly chaotic. Therefore, we introduce a novel dynamic learning strategy to optimize the agent election mechanism and adaptive collaboration network. By refining both the election and review processes, our framework seeks to discover the most effective communication patterns among agents, thereby improving geo-localization performance through collaborative reasoning while minimizing unnecessary discussions. In summary, our contributions are demonstrated as follows:

- We propose a novel framework, smileGeo, that adaptively integrates both the inherent and retrieved knowledge, along with the reasoning capabilities of LVLMs, through structured discussions for visual geo-localization tasks.
- We introduce a dynamic learning strategy to identify the most effective communication patterns among LVLM agents, enhancing both effectiveness and efficiency.
- We conducted experiments on two open-source datasets. To address the issue of numerous images (food, furniture, *etc.*) in these datasets that could not be localized, we also constructed a new dataset for further evaluation. Extensive experimentation demonstrates that smileGeo achieves competitive performance compared to state-of-the-art methods.

2 Methodology

In this section, we first present the overall framework and then introduce each part of smileGeo in detail for geo-localization tasks.

2.1 Model Overview

In this paper, we denote the social network of LVLM agents by $\mathcal{G}$, where $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$. $\mathcal{V}$ stands for the agent set and $\mathcal{E}$ presents the set of edges. Each agent $v_i \in \mathcal{V}, i \in [N]$ is an LVLM, which is pre-trained by massive vision-language data and can infer the possible location Y of a given image X. Besides, each edge $e_{ij} \in \mathcal{E}, i, j \in [N]$ is the connection weighted by the improvement effect of agent v_i to agent v_j via discussion regarding the geo-localization performance.

As illustrated in Figure 2, smileGeo contains the process of the review mechanism in agent discussions along with a dynamic learning strategy of agent social networks. In this framework, the review mechanism in agent discussions is a 3-stage anonymous collaboration approach to allow retrieval-augmented LVLM agents to reach a consensus via discussion:

In the first stage, for a given image X, our proposed framework, smileGeo, elects K suitable agents as answer agents by agent election probability $\boldsymbol{Lst} := \{P_{v_1}, P_{v_2}, \cdots\}$, where P_{v_i} is the probability of selecting agent v_i.

In the second stage, these answer agents respectively select R review agents by the adaptive collaboration social network $\boldsymbol{A}$, which is the adjacency matrix generated from $\mathcal{G}$ dynamically, to refine their answer via discussion.

In the third stage, our proposed framework facilitates consensus among all agents through open discussion to reach a final answer. Both $\boldsymbol{Lst}$ and $\boldsymbol{A}$ are analyzed from the given image X, allowing our framework to minimize unnecessary discussions, thereby significantly enhancing its efficiency while maintaining its accuracy. Moreover, the multi-stage discussion facilitates communication among agents, maximizing the integration of their knowledge and reasoning abilities to generate an accurate response Y.

To get $\boldsymbol{Lst}$ and $\boldsymbol{A}$, we specifically design a dynamic learning module in parallel with the aforementioned 3-stage discussion, which initially deploys the encoder component of a pre-trained image variational autoencoder (VAE) to extract features from the given image X. The extracted features, combined with learnable agent embeddings $\boldsymbol{Emb}$, are employed through a GNN-based module to determine the suitability of agents *w.r.t,* $\boldsymbol{Lst}$ for agent discussions and predict suitable agent collaboration connections $\boldsymbol{A}$ according to the given geo-localization query.

2.2 Review Mechanism in Agent Discussions

LVLM have demonstrated remarkable capabilities in complicated tasks and some pioneering works have further proven that the performances can be further enhanced by ensembling multiple LVLM agents [7, 21, 26, 39, 50]. Thus, to improve the geo-localization capability of LVLMs, we propose a cooperation framework to effectively integrate the diverse knowledge (including external retrieved information) and reasoning abilities of multiple LVLMs. Inspired by the fact that community review mechanisms can improve the quality of manuscripts, an iterative 3-stage anonymous reviewing mechanism is proposed for helping agents share knowledge and reasoning capability with each other through their collaboration social network: i) answer agent election & answering, ii) review agent selection & reviewing, and iii) final answer conclusion.

Stage 1: Answer Agent Election & Answering
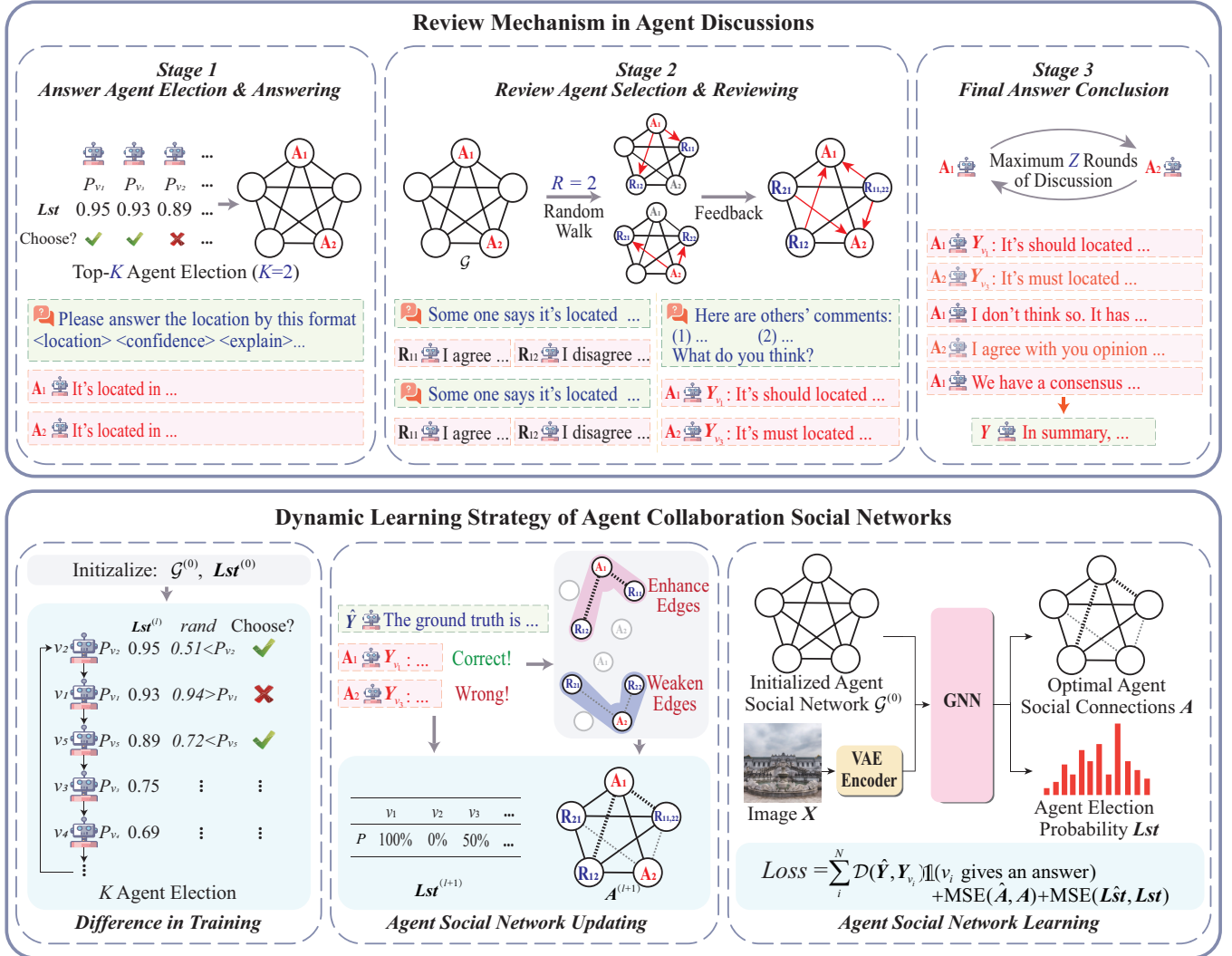
Figure 2: The framework overview of smileGeo. It contains the process of review mechanism in agent discussions along with a dynamic learning strategy of agent collaboration social networks. The first part deploys a review mechanism for LVLMs to discuss and share their knowledge anonymously, which could enhance the overall performance of geo-localization tasks. The second one mainly utilizes the GNN-based learning module to improve efficiency by reducing unnecessary discussions among agents while showing the process of updating the agent collaboration social network during the training process.

Initially, we select K agents with the highest agent election probabilities Lst as answer agents and let them geo-localize independently as the preliminary step for further discussion. By initiating the discussion with a limited number of agents, we aim to reduce potential chaos and maintain the efficiency of our framework as the number of participating agents increases.

After the answer agents are elected, we send the image X to all answer agents and let them give the primary analysis. Each answer must contain three parts: one location (city, country, and so on), one confidence (a percentage number), and a detailed explanation. Besides, if an agent is unable to interpret the given image, it is permitted to utilize a combination of network search and chain-of-thought reasoning [44] to gather additional relevant information.

Stage 2: Review Agent Selection & Reviewing

In this stage, for each answer agent, we choose R retrieval-augmented review agents by performing a transfer-probability-based random walk on the agent collaboration social network $\mathcal{G}$ for answer reviewing. The transfer probability $p(v_i, v_j)$ from node v_i to node v_j can be calculated as follows:

$$p(v_i, v_j) = \begin{cases} \dfrac{A_{ij}}{\sum_{k \in \mathcal{N}(v_i)} A_{ik}}, & \text{if } e_{ij} \in \mathcal{E}, \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $\mathcal{N}(v_i)$ is the 1-hop neighbor node set of node v_i.

For each selected review agent, it reviews the results as well as the explanations generated by the corresponding answer agent and gives its own comments. After that, each answer agent would summarize their preliminary analysis and the feedback from all of

its review agents to get the final answer, which must include three parts as well: one location, one confidence, and an explanation.

Stage 3: Final Answer Conclusion

In the previous stage, each answer agent produces a refined result based on feedback. When $K > 1$ in Stage 1, the proposed framework generates multiple independent results, which may not be consistent. However, we aim to provide a definitive answer rather than multiple options for people to choose from. To address this, we allow up to Z rounds of free discussion among those answer agents to reach a unified answer:

First, we maintain a global dialog history list, $diag$, recording all replies agents respond. In addition, discussions are executed asynchronously, which means that any answer agent can always reply based on the latest $diag$, and replies would be added to the end of $diag$ as soon as they are posted. Each answer agent is allowed to speak only once in each discussion round, and after Z rounds of free discussion, we determine the final result using a minority-majority approach, *i.e.,* we choose the reply with the most agreement as the final conclusion. If all agents reach a consensus (or when the conclusion of the majority of agents is clearly dominant), we early stop this stage and adopt the consensus answer as the final answer. If none of any consensus is reached, we only select the reply of the first answer agent elected from Stage 1 as the final result.

2.3 Dynamic Learning Strategy of Agent Collaboration Social Networks

In our framework, choosing the appropriate answer agents and review agents for knowledge sharing and discussion is vital to its effectiveness and efficiency. Therefore, we propose a dynamic learning strategy to optimize them. Specifically, for each training sample, *i.e.,* a geo-tagged image, we would first estimate the optimal answer agent election probability $\hat{Lst}$ and the optimal collaboration social network of agent $\hat{\mathcal{G}}$ by its actual location. Then we train an attention-based graph neural network, which aims to predict Lst and $\mathcal{G}$, by such estimated ground truth.

To estimate the optimal $\hat{Lst}$ and $\hat{A}$ for agents to geo-localize image X, we first initialize the agent social network $\mathcal{G}^{(0)}$ by a fully connected graph with the agent set $\mathcal{V}$. Besides, we initialize the agent election probability $Lst^{(0)} = [0.5, 0.5, \cdots]$, with all agents having 50% probability of being chose as answer agents.

Then, we iteratively conduct our 3-stage discussion framework to get the prediction answer. $Lst^{(l)}$ and $\mathcal{G}^{(l)}$ is updated at the end of each round $l \in L$ by comparing the answers $Y_{v_i}^{(l)}$ from each answer agent with the ground truth $\hat{Y}$.

After L rounds of agent discussions, the updated agent election probability for an image X, $\hat{Lst} := Lst^{(L)}(X) = [P_{v_1}^{(L)}, P_{v_2}^{(L)}, \cdots, P_{v_N}^{(L)}]$, determines whether an agent v_i gives the correct/wrong answers $Y_{vi}^{(L)}$ by comparing it with the ground truth $\hat{Y}$. Here, the definition of $P_{v_i}^{(l)}$ of agent v_i at round l is as follows:

$$P_{v_i}^{(l)} := \begin{cases} 0, & \text{if } \mathcal{D}(\hat{Y}, Y_{v_i}^{(l)}) > th, \\ 1, & \text{if } \mathcal{D}(\hat{Y}, Y_{v_i}^{(l)}) \le th, \\ \frac{1}{2}, & \text{if } v_i \text{ did not participate in the discussion,} \end{cases} \quad (2)$$

where th is a pre-defined threshold for determining whether the predicted location is close enough to the actual location. In the distance function $\mathcal{D}(\cdot)$, we first deploy geocoding to convert natural language into location intervals in a Web Mercator coordinate system (WGS84) by utilizing OSM APIs, and then compute the shortest distance between two two location intervals.

Please note that, rather than electing the top-K answer agents in each round, we choose each agent with probability P_{v_i} during the training period to ensure that every agent has the opportunity to participate in the discussion for more accurate estimation, as shown at the left part of the dynamic learning strategy module of agent collaboration social networks in Figure 2.

In addition, the agent collaboration social network would also be updated by comparing the actual location with the generated answer of each answer agent at the same time. For l-th round, we strengthen the link between the correctly answered agent and the corresponding review agents while weakening the link between the incorrectly answered agent and the corresponding review agents:

$$\hat{A}_{ij} := A_{ij}^{(l)}(X) = \begin{cases} \frac{tt+1}{2tt} A_{ij}^{(l-1)}(X), & \text{if } v_i \text{ answers correctly,} \\ \frac{2tt-1}{2tt} A_{ij}^{(l-1)}(X), & \text{if } v_i \text{ answers incorrectly,} \end{cases} \quad (3)$$

where $A_{ij}^{(l-1)}(X)$ is the weight of the connection between answer agent v_i and review agent v_j at round $l-1$ when geo-locating image X, $A_{ij}^{(0)}(X) = 1, i \ne j, A_{ii}^{(0)}(X) = 0, i, j \in [N]$, tt is the number of consecutive times an agent has answered correctly, which is used to attenuate the connection weights when updating them, preventing the performance of an agent on a certain portion of the continuous dataset from interfering with the model's evaluation of the current agent's performance on the entire dataset.

Then, we try to learn an attention-based graph neural network to predict the corresponding optimal agent election probability $Lst = h(X, \mathcal{G}|\Theta)$ and the optimal agent collaboration connections $A = f(X, \mathcal{V}|\Theta)$:

$$\begin{aligned} A &= \text{Att}_{\text{GNN}}(Fea, Fea, 1) \\ &= \text{softmax}\left(\frac{Fea \cdot Fea^\top}{\sqrt{d_k}}\right) 1, \\ Lst &= \sigma'\left(\text{Linear}\left(\text{Flatten}\left(\sigma\left(A \cdot Fea \cdot W\right)\right)\right)\right), \\ Fea &= \text{Linear}\left(Emb + \text{VAE}_{\text{Enc}}(X)\right), \end{aligned} \quad (4)$$

where $W, Emb \in \Theta$ are two learnable parameters, $Emb := [Emb_{v_1}, Emb_{v_2}, \cdots]^\top$ is the agent embedding and W is the weight matrix, $\sigma(\cdot)$ is the LeakyReLU function, $\sigma'(\cdot)$ is the Sigmoid function, $\text{VAE}_{\text{Enc}}(\cdot)$ is the encoder of the image VAE that compresses and maps the image data into the latent space. It is used to align the image features with the agent embedding, and d_k is the dimension of the Fea. Our learning target can be formalized as:

$$\begin{aligned} \arg\min_{\Theta} \sum_i^N \mathcal{D}(\hat{Y}, Y_{v_i}) \mathbb{1}(v_i \text{ gives an answer}) \\ + \text{MSE}(\hat{Lst}, Lst) + \text{MSE}(\hat{A}, A), \end{aligned} \quad (5)$$

where $\mathcal{D}(\cdot)$ denotes the distance between the places an LVLM agent answered and the ground truth, $\mathbb{1}(\cdot)$ is the indicator function, $Y_{v_i} := Y_{v_i}^{(L)} = g_{v_i}(X, Y_{v_j}^{(L-1)})$, $g_{v_i}(\cdot)$ represent the LVLM agent v_i with fixed parameters and $Y_{v_i}^{(0)} = g_{v_i}(X)$ is the answer that LVLM agent v_i generates at the initial stage of discussion.

3 Experiments

To evaluate the performance of our framework, we conducted experiments on the real-world dataset that was gathered from the Internet to answer the following research questions:

• **RQ1**: Can the proposed framework, smileGeo, outperform state-of-the-art methods in open-ended geo-localization tasks?

• **RQ2**: Are LVLMs with diverse knowledge and reasoning abilities more suitable for building a collaborative social network of agents?

• **RQ3**: How efficient is smileGeocompared to other baselines?

• **RQ4**: How does the setting of different hyperparameters affect the performance of smileGeo?

3.1 Experiment Setup

Datasets. In this paper, we first evaluate the proposed geo-localization framework, smileGeo, on the two open-source datasets: IM2GPS3K[1] and YFCC4K[2]:

The IM2GPS3K dataset is a widely used benchmark for visual geo-localization. It consists of 3,000 images from the Flickr photo-sharing platform, which are tagged with precise geographical coordinates. The dataset covers a diverse range of locations globally, including urban, rural, and natural environments, making it an ideal testbed for assessing the generalization capabilities of geo-localization models. The images in IM2GPS3K are drawn from various categories, such as landscapes, cityscapes, and landmarks, providing a challenging and diverse set of visual cues that the models must recognize to predict the geographical location accurately.

The YFCC4K dataset is a subset of the larger Yahoo Flickr Creative Commons 100 Million (YFCC100M) dataset, which contains over 100 million Flickr images. The YFCC4K subset includes 4,536 images with corresponding geographical tags, selected to represent a broad geographic distribution and visual diversity. Like IM2GPS3K, YFCC4K includes a variety of image types, such as natural landscapes, urban settings, and iconic landmarks. This diversity allows for comprehensive evaluation of geo-localization models across different environments and scales, from global-level to city-level localization tasks.

Noting that the labels in the above two datasets are latitude-and-longitude-based GPS points rather than natural language, we use geo-reverse encoding technology to map GPS points into detailed and structured addresses for evaluation.

In addition, We have newly constructed a geo-localization dataset named GeoGlobe[3]. It contains a variety of man-made landmarks or natural attractions from nearly 150 countries with different cultural and regional styles. The diversity and richness of GeoGlobe allow us to evaluate the performance of different models more accurately. The images in this dataset are copyright-free images obtained from the Internet via a crawler. We divide the images into two main

categories: man-made landmarks and natural attractions. Then, we filter out the data samples that could clearly identify the locations of the landmarks or attractions in the images. As a result, we filter out nearly three hundred thousand data samples, and please refer to Table 1 and Figure 3 for details. Due to the fact that a large number of natural attractions in different geographical regions with high similarity are cleaned, the magnitude of the data related to natural attractions in this dataset is smaller than that of man-made attractions.
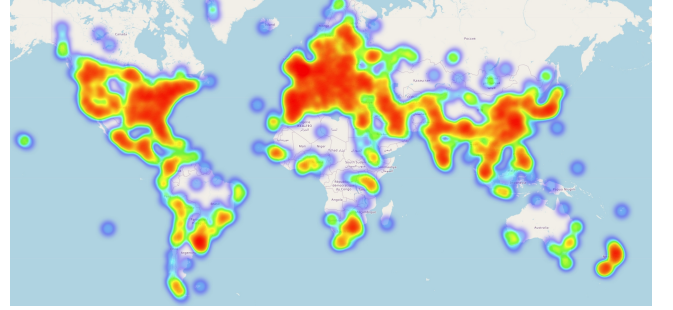

Figure 3: The data distribution around the world.

Table 1: Statistics of the dataset GeoGlobe.

	Images	Cities	Countries	Attractions
ManMade	252,375	2,313	143	10,492
Natural	39,996	1,044	97	1,849

For an open-world geo-localization task, the relationship between the training and test samples in the experiment could greatly affect the results. We label the locations of training samples as $\mathcal{Z}_{\text{train}}$, and the ones of test sample set as $\mathcal{Z}_{\text{test}}$, and use two metrics, *coverage* as well as *consistency*, to portray this relationship:

$$
\begin{aligned}
coverage &= \frac{\mathcal{Z}_{\text{train}} \cap \mathcal{Z}_{\text{test}}}{\mathcal{Z}_{\text{train}}} \times 100\%, \\
consistency &= \frac{\mathcal{Z}_{\text{train}} \cap \mathcal{Z}_{\text{test}}}{\mathcal{Z}_{\text{test}}} \times 100\%.
\end{aligned}
\tag{6}
$$

As for the samples in this paper, $coverage \approx 4.6564\%$, and $con-sistency \approx 33.2957\%$.

Baselines. In this work, we mainly compare the proposed framework with different geolocation retrieval systems and deep learning methods: including NetVLAD [1], GeM [34], CosPlace [2], Translocator [33], and GeoCLIP [40]. For image retrieval systems, we set the whole training dataset as the geo-tagged image database and only query images in the test dataset for systems to answer. Besides, we select both open-source and close-source LVLMs with different scales pretrained by various datasets as agents in the proposed framework. As for the open-source LVLMs, we utilize several open-source fine-tuned LVLMs: Infi-MM[4], Qwen-VL[5], vip−llava−7b&13b[6], llava−1.5−7b−base&mistral&vicuna[7], llava−1.6−7b&13b&34b−mistral&vicuna[8], CogVLM[9]. As for the closed-source LVLMs, we chose the models provided by three of the most

famous companies in the world: Claude–3–opus[10], GPT–4o–mini[11], and Gemini–1.5–pro [12].

Implementation Details. In all experiments, we employ a variety of LVLMs, encompassing both open-source and closed-source models, to be agents in smileGeo. Unless specified otherwise, zero-shot prompting is applied. Besides, 99% of images from each dataset are randomly chosen as training samples. For the open-world geo-localization problem, we constructed the test dataset that is entirely independent of each training dataset. Additionally, approximately 50%-60% of the test samples consist of images from distinct locations with no overlap with the training data. Each open-source LVLM is deployed on a dedicated A800 (80G) GPU server with 200GB memory. As for each closed-source LVLM, we cost billions of tokens by calling APIs as specified by the official website. Additionally, we use a dedicated line to directly access the API from a server in a US data centre to minimize communication delays. To circumvent the rate limits on closed-source API calls, we employ hundreds of different accounts to initiate queries simultaneously. To avoid the context length issue that occurs in some LVLMs, we truncate the context before submitting it to the agent for questions based on the maximum number of tokens that each agent supports. Furthermore, since images are token-intensive, we only retain the latest response when facilitating discussions between different agents. More details about the deployment of smileGeo and the settings of related hyperparameters can be found in **Appendix B**.

Evaluation Metrics. We use *Accuracy* (*Acc*) to evaluate the performance: $Acc = \frac{N_{\text{correct}}}{N_{\text{total}}}$, where N_{correct} is the number of samples that the proposed framework correctly geo-localizes, and N_{total} refers to the total number of testing samples.

3.2 Performance Comparison

Table 2 presents a comparison between our proposed framework and all baseline approaches. Our framework consistently outperforms all other methods. This superior performance can be attributed to the limitations of traditional image retrieval techniques, which rely heavily on rich geo-tagged image databases and exhibit constrained reasoning capabilities. In contrast, our method can effectively analyze and integrate results retrieved from the Internet, enabling it to calculate more accurate geo-localization outcomes. Moreover, over half of the images in our test dataset are new and localized in areas completely different from those in the training dataset. This underscores the shortcomings of conventional database-based retrieval systems, particularly due to the inherent limitations of geo-tagged image databases, and demonstrates the effectiveness of our framework in addressing open-world geo-localization tasks.

It is also worth noting that the YFCC4K and IM2GPS3K datasets do not apply artificial filtering to images, resulting in ambiguous content with minimal geographical cues, such as food photos and portraits. Comparing the model performance across different datasets, we observe that it performs best on the GeoGlobe dataset.

Table 2: Comparison with baselines (*Acc*, %).

Model	IM2GPS3K	YFCC4K	GeoGlobe (Natural)	GeoGlobe (ManMade)
NetVLAD	16.6333	7.4956	26.4339	29.0004
GeM	14.5000	6.5256	23.1920	25.4050
CosPlace	17.6667	7.9586	28.1796	30.2647
Translocator	31.1000	13.4039	26.1845	34.2157
GeoCLIP	34.4667	15.1675	38.1546	45.9107
smileGeo	**47.7667**	**21.5168**	**76.0599**	**85.4603**

Bold indicates the statistically significant improvements (*i.e.,* two-sided t-test with $p < 0.05$) over the best baseline.

Table 3: Results of different single LVLM baselines (*Acc*, %).

Model	IM2GPS3K	YFCC4K	GeoGlobe (Natural)	GeoGlobe (ManMade)
Infi-MM	14.7000	6.5256	19.2020	21.4145
Qwen-VL	32.4667	14.3959	42.3940	37.4556
vip-llava-13b	15.7667	6.9885	20.6983	15.4089
vip-llava-7b	38.7000	17.8792	31.9202	56.4994
llava-1.5-7b	30.2000	13.8889	27.1820	47.2145
llava-1.6-7b-mistral	3.1000	1.6314	0.7481	2.1731
llava-1.6-7b-vicuna	9.4667	4.4092	6.9825	15.8831
llava-1.6-13b	17.8333	8.3333	12.2195	28.2497
llava-1.6-34b	44.2667	20.8113	52.8678	77.2027
CogVLM	6.7333	3.0644	7.7307	10.3516
Claude-3-opus	23.8333	12.3457	33.1671	40.6954
GPT-4o-mini	45.0000	18.3422	62.0948	84.5911
Gemini-1.5-pro	47.3667	19.9956	62.3441	82.8131
smileGeo	**47.7667**	**21.5168**	**76.0599**	**85.4603**

Bold indicates the statistically significant improvements (*i.e.,* two-sided t-test with $p < 0.05$) over the best baseline.

3.3 Ablation Study

Effect of LVLM Discussion. In this section, we aim to verify that the observed performance improvements stem from the collaborative discussion among multiple LVLM agents rather than from a single LVLM agent. To achieve this, we employ the chain-of-thought (CoT) method [44] to evaluate each agent individually within our framework, and the results are presented in Table 3. We can find that smileGeo outperforms any individual LVLM across all datasets, including the most advanced closed-source models such as GPT-4o-mini and Gemini-1.5-pro. Furthermore, the varying performances of different single LVLMs on different datasets highlight that each agent possesses distinct knowledge and reasoning capabilities. Our proposed framework facilitates effective information exchange among agents, thereby enhancing the reasoning abilities of the models for diverse geo-localization tasks.
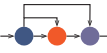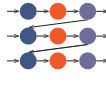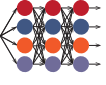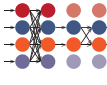
Power of Internet-enabled Information Retrieval. To demonstrate the benefits of allowing every LVLM agent to retrieve information from the Internet for enhancing the knowledge of the model, as well as the potential for multiple LVLMs to use the retrieved data for further reasoning, we designed an experiment as shown in Figure 4. Figure 4(a) illustrates the results without any retrieval, while Figure 4(b) displays the results when models incorporate the retrieved information. We compare the proposed model with two

---

[10]https://anthropic.com/
[11]https://openai.com/
[12]https://gemini.google.com/

Table 4: Results of different LVLM agent frameworks testing on GeoGlobe (ManMade).

Framework	LLM-Blender	PHP	Reflexion	LLM Debate	DyLAN	smileGeo
Structure						
$Acc \uparrow$	75.3457%	82.3785%	84.1960%	77.0051%	84.8281%	**85.4603%**
$Tks \downarrow$	23,662	154,520	109,524	260,756	159,320	**17,145**

'Acc' stands for the accuracy of the framework;
'Tks' means the average tokens a framework costs per query (including image tokens).
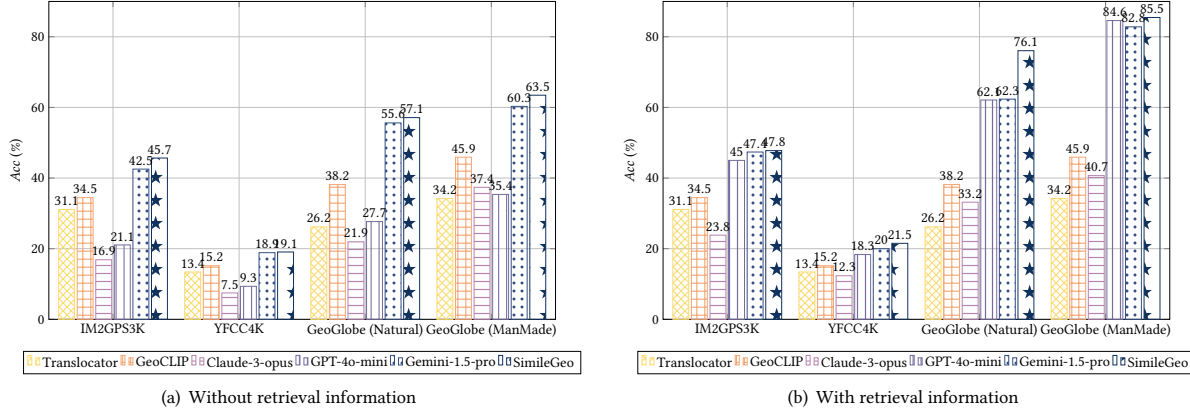

(a) Without retrieval information

(b) With retrieval information

Figure 4: The impact of Internet-enabled information retrieval.

types of methods: image similarity-based retrieval approaches and advanced closed-source large models. The results indicate that image similarity-based methods, such as Translocator and GeoCLIP, struggle to effectively leverage Internet-retrieved information, leading to minimal changes in accuracy. In contrast, closed-source large models show moderate performance improvements when aided by additional retrieval data. Notably, our method consistently outperforms all baselines across datasets, with a significant improvement of over 20% in the GeoGlobe (natural) dataset for smileGeo. This demonstrates that our proposed framework, with its robust reasoning capabilities, can fully exploit retrieved information to enhance reasoning and model accuracy.

Different LVLM Agent Structures. We experiment with multi-agent collaborative frameworks with the same LVLM agents but different structures, including LLM-Blender [21], PHP [50], Reflexion [39], LLM Debate [7], and DyLAN [26]. The comparative results across various LVLM agent frameworks are presented in Table 4. It is evident that the majority of LVLM agent frameworks surpass individual LVLMs in terms of geo-localization accuracy in Table 3. This improvement can primarily be attributed to the ability to integrate knowledge from multiple LVLM agents, thereby enhancing the overall precision of these frameworks. However, LLM-Blender and LLM Debate exhibit lower accuracy due to statements of some agents misleading others during discussions, which impedes the generation of correct outcomes. Our framework, smileGeo, guarantees the highest accuracy while being able to accomplish the geo-localization task with the lowest token costs. The average number of tokens our framework spent per query is 17,145, and it is less

than the computational overhead of LLM-Blender (23,662), which has the simplest agent framework structure but the lowest accuracy among all baselines. This is mainly due to a 'small' GNN-based dynamic learning model being deployed for agent selection stages and significantly reducing unnecessary discussions among agents.

3.4 Efficiency Study

As illustrated in Table 4, the proposed model achieves higher accuracy while consuming fewer tokens. To further demonstrate this efficiency, we introduce a more intuitive metric, *Response Time* (*RT*), which reflects the overall performance of smileGeo.

Our framework adapts to different application scenarios by utilizing the variable Z, as defined in stage 3 of the methodology section, to strike a balance between model accuracy and efficiency. For example, when analyzing historical mobility patterns, a higher Z value can be used since this task is less sensitive to response

Table 5: Model efficiency testing on GeoGlobe (ManMade).

Z	RT_{Avg} (ms)	RT_{Med} (ms)	Acc (%)
0	473	473	75.2667
1	608	607	77.8349
3	766	754	81.5889
5	1,127	1,028	85.4603
10	1,897	1,123	86.2110
15	10,268	1,139	86.6456
20	22,176	1,289	87.0407
50	24,056	1,454	87.5543

(a) Calling the same LVLM API

(b) Different LVLM backbones
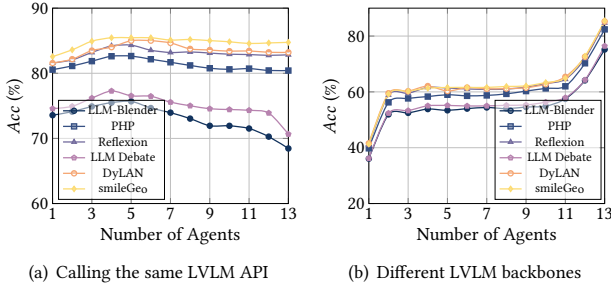
Figure 5: Results of model performance in relation to the number of agents.

Figure 6: Results under different K and R.

time but requires a higher accuracy. Conversely, for tasks such as recommending tourist attractions based on dynamic user posts, a lower Z value is preferable to prioritize faster responses.

Table 5 reports the average and median of response times (RT_{Avg}, RT_{Med}), as well as the model accuracy (Acc). In general, as the maximum allowable discussion round Z increases, the response time of smileGeo also increases. However, thanks to implementing our dynamic learning model with an early stopping strategy, this growth remains nearly linear rather than exponential.

Then, the comparison of RT_{Avg} and RT_{Med} shows that only a small fraction of image queries require multiple discussion rounds. These outliers, with significantly longer response times, skew the average upwards. From the median response time, we observe that at least 50% of the test samples receive a response within 2 seconds, an acceptable performance level for most applications.

3.5 Parameter Analysis

Number of Agents. We further demonstrate the relationships between the number of agents and the framework performance. We conduct experiments in two ways: (i) by calling the same closed-source LVLM API (Here, we use Gemini-1.5-pro because it performs best without the help of the Internet) under different prompts (*e.g.,* You are good at recognizing natural attractions; You're a traveler around Europe) to simulate different agents, and (ii) by using different LVLM backbones to represent distinct agents. The results are shown in Figure 5.

As illustrated in Figure 5(a), the framework achieves optimal accuracy with 4 or 5 agents. Beyond this number, the framework's performance begins to deteriorate. This shows that using the same backbone model with the fixed knowledge and reasoning capabilities as different agents has limited improvement in the accuracy of the framework. Despite this decline, the performance of frameworks other than LLM-Blender and LLM Debate remains superior to that of a single agent. LLM-Blender and LLM Debate, however, have a significant decrease in model accuracy when the number of agents exceeds 11. This is mainly because both of them involve all LVLMs in every discussion, which suffers from excessive repetitive and redundant discussions. Figure 5(b) reveals that the accuracy of the framework improves with the incorporation of more LVLM backbones, indicating that the diversity of LVLM agents can enhance the quality of discussions.
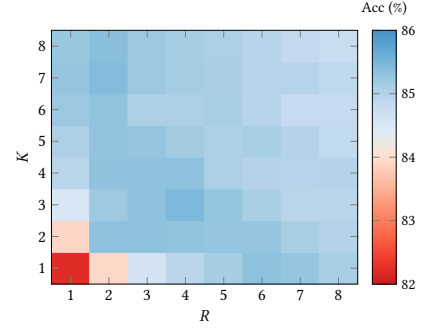
Hyperparameter K & R. There are two hyperparameters, K and R, that need to be pre-defined in the proposed framework: K is the number of agents (answer agents) that respond in each round of discussion, and R is the number of agents (review agents) used to review answers from answer agents. Therefore, we conduct experiments under different combinations of $K \in [1, 8]$ and $R \in [1, 8]$, as shown in Figure 6. The results indicate that optimal performance can be achieved with relatively small values of K or R. However, the computational cost, measured in tokens, increases exponentially with higher values of K and R. To ensure the accuracy of the smileGeo while reducing the calculation cost as much as possible, we set both K and R equal to 2 in this paper.

3.6 Case Study

To further illustrate the superiority of our proposed framework, we provide detailed examples. Additional information about the case study is presented in **Appendix C**.

4 Discussion

As shown in Figure 7(a), the data retrieved from the internet frequently includes a significant amount of irrelevant advertising information. This extraneous content hampers the model's accuracy in image localization. Figure 7(b) further reveals that, after human workers manually filtered out the advertising content, 80% of the retrieval results were unrelated to the queried image. This underscores the necessity of enhancing the model's resilience to such interference, which will be a major focus of our future research.

Another critical challenge is the computational efficiency, which is a common limitation of LVLMs. Currently, a range of methods exists, such as model distillation and compression, which can enhance computational efficiency. For instance, in our current implementation, we used high-performance servers located in U.S. data centres to support the system and deployed smaller models, like GPT-4o-mini, to reduce response times. According to our experiments, GPT-4o-mini demonstrated substantially improved response times compared to previous models, such as GPT-4V. We anticipate that ongoing advancements in large-model technology will continue to address these efficiency concerns. In the future, we plan to incorporate distillation-based models as agents within our framework, aiming to enhance both computational performance and robustness. It is expected to improve system response time and

(a) List of original Internet search results   (b) List of results after manual filtering
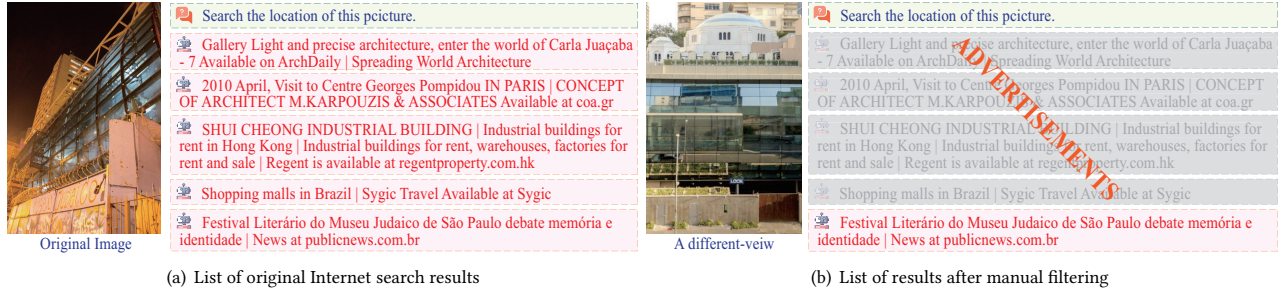
Figure 7: A case study demonstrating retrieval results contains a lot of misleading information.

resilience against interference further by enabling faster decision-making and better adaptability to noisy or irrelevant information.

5 Related Work

Visual Geo-localization. Recent research in visual geo-localization, commonly referred to as geo-tagging, primarily focuses on developing image retrieval systems to address this challenge [1, 12, 22, 25, 31, 43]. These systems utilize learned embeddings generated by a feature extraction backbone, which includes an aggregation or pooling mechanism [14, 17, 32, 34]. However, the applicability of these retrieval systems to globally geo-localize landmarks or natural attractions is often limited by the constraints of the available database knowledge and the restrictions imposed by national or regional geo-data protection laws. Alternatively, some studies treat visual geo-localization as a classification problem [19, 24, 29, 37]. These approaches posit that two images from the same geographical region, despite depicting different scenes, typically share common semantic features. Practically, these methods organize the geographical area into discrete cells and categorize the image database accordingly. This cell-based categorization facilitates scaling the problem globally, provided the number of categories remains manageable. However, while the number of countries globally remains relatively constant, accurately enumerating cities in real-time at a global scale is challenging due to frequent administrative changes, such as city reorganizations or mergers, which reflect shifts in national policies. Additionally, in the context of globalization, this strategy has inherent limitations. The recent advent of LVLMs offers promising compensatory mechanisms for the deficiencies observed in traditional geo-localization methodologies, making the exploration of LVLM-based approaches significantly relevant in current research.

Multi-agent Framework for LLM/LVLMs. LLM/LVLM agents have demonstrated the potential to act like human [3, 30, 35], and a large number of studies have focused on developing robust architectures for collaborative LLM/LVLM agents [7, 21, 26, 39, 50]. These architectures enable each LLM/LVLM agent that endows with unique capabilities to engage in debates or discussions. For instance, [21] proposes an approach to aggregate multiple LLM/LVLM responses by generating candidate responses from various LLM/LVLM in a single round and employing pairwise ranking to synthesize the most effective response. While some studies [21] utilize a static architecture potentially limiting the performance and generalization of LLM/LVLM, others like [26] have implemented dynamic interaction architectures that adjust according to the query and

incorporate user feedback. Recent advancements also demonstrate the augmentation of LLM/LVLM as autonomous agents capable of utilizing external tools to address challenges in interactive settings. These techniques include retrieval augmentation [18, 38, 47], mathematical tools [27, 36, 47], and code interpreters [10, 42]. With these capabilities, LLM/LVLMs are well-suited for various tasks, especially for geo-localization. However, most LLM/LVLM agent frameworks mandate participation from all agents in at least one interaction round, leading to significant computational overhead. To address this issue, our framework introduces a dynamic learning strategy electing only a small number of agents to geo-localize different images, which significantly enhances the efficiency of LLM/LVLM agents by reducing unnecessary interactions.

6 Conclusion

This work introduces a novel LVLM agent framework, smileGeo, specifically designed for geo-localization tasks. Inspired by the review mechanism, it integrates various LVLMs to discuss anonymously and geo-localize images worldwide. Additionally, we have developed a dynamic learning strategy for agent collaboration social networks, electing appropriate agents to geo-localize each image with different characteristics. This enhancement reduces the computational burden associated with collaborative discussions among LVLM agents. Moreover, we have constructed a geo-localization dataset called GeoGlobe to evaluate the proposed framework better. Overall, smileGeo demonstrates significant improvements in geo-localization tasks, achieving superior performance with lower computational costs compared to state-of-the-art baselines.

Looking ahead, we aim to expand the capabilities of smileGeo to incorporate more powerful external tools beyond just web searching. Additionally, we plan to explore extending its application to more complex scenarios, such as high-precision global positioning, laying the cornerstone for exploring LVLM agent collaboration to handle different complex open-world tasks efficiently.

Acknowledgments

References

[1] Relja Arandjelovic, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. 2018. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* (2018), 1437–1451.

[2] Gabriele Moreno Berton, Carlo Masone, and Barbara Caputo. 2022. Rethinking Visual Geo-localization for Large-Scale Applications. In *Proc. of CVPR.* 4868–4878.

[3] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR* (2023).

[4] Mark Campbell and Matt Wheeler. 2006. A vision based geolocation tracking system for uav's. In *AIAA Guidance, Navigation, and Control Conference and Exhibit.* 6246.

[5] Zetao Chen, Lingqiao Liu, Inkyu Sa, Zongyuan Ge, and Margarita Chli. 2018. Learning Context Flexible Attention Model for Long-Term Visual Place Recognition. *IEEE Robotics Autom. Lett.* (2018), 4015–4022.

[6] Fang Deng, Lele Zhang, Feng Gao, Huangbin Qiu, Xin Gao, and Jie Chen. 2020. Long-Range Binocular Vision Target Geolocation Using Handheld Electronic Devices in Outdoor Environment. *IEEE Trans. Image Process.* (2020), 5531–5541.

[7] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *CoRR* (2023).

[8] Moataz Medhat ElQadi, Myroslava Lesiv, Adrian G. Dyer, and Alan Dorin. 2020. Computer vision-enhanced selection of geo-tagged photos on social network sites for land cover classification. *Environ. Model. Softw.* (2020), 104696.

[9] Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. 2024. Large Language Model-based Human-Agent Collaboration for Complex Task Solving. *arXiv preprint arXiv:2402.12914* (2024).

[10] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided Language Models. In *Proc. of ICML.* 10764–10799.

[11] Sourav Garg, Niko Sünderhauf, and Michael Milford. 2022. Semantic-geometric visual place recognition: a new perspective for reconciling opposing views. *Int. J. Robotics Res.* (2022), 573–598.

[12] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. 2020. Self-supervising Fine-Grained Region Similarities for Large-Scale Image Localization. In *Proc. of ECCV.* 369–386.

[13] Xiao Han, Zijian Zhang, Xiangyu Zhao, Yuanshao Zhu, Guojiang Shen, Xiangjie Kong, Xuetao Wei, Liqiang Nie, and Jieping Ye. 2025. GARLIC: GPT-Augmented Reinforcement Learning with Intelligent Control for Vehicle Dispatching. In *Proc. of AAAI.* 255–262.

[14] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. 2021. Patch-NetVLAD: Multi-Scale Fusion of Locally-Global Descriptors for Place Recognition. In *Proc. of CVPR.* 14141–14152.

[15] Stephen Hausler, Adam Jacobson, and Michael Milford. 2019. Multi-Process Fusion: Visual Place Recognition Using Multiple Image Processing Methods. *IEEE Robotics Autom. Lett.* (2019), 1924–1931.

[16] Binxuan Huang and Kathleen M. Carley. 2019. A large-scale empirical study of geotagging behavior on Twitter. In *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019.* 365–373.

[17] Sarah Ibrahimi, Nanne van Noord, Tim Alpherts, and Marcel Worring. 2021. Inside Out Visual Place Recognition. In *Proc. of BMVC.* 362.

[18] Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *J. Mach. Learn. Res.* (2023), 251:1–251:43.

[19] Mike Izbicki, Evangelos E. Papalexakis, and Vassilis J. Tsotras. 2019. Exploiting the Earth's Spherical Geometry to Geolocate Images. In *Proc. of KDD.* 3–19.

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

[21] Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. In *Proc. of ACL.* 14165–14178.

[22] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. 2017. Learned contextual feature reweighting for image geo-localization. In *Proc. of CVPR.* 2136–2145.

[23] Ahmad Khaliq, Shoaib Ehsan, Zetao Chen, Michael Milford, and Klaus D. McDonald-Maier. 2020. A Holistic Visual Place Recognition Approach Using Lightweight CNNs for Significant ViewPoint and Appearance Changes. *IEEE Trans. Robotics* (2020), 561–569.

[24] Giorgos Kordopatis-Zilos, Panagiotis Galopoulos, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2021. Leveraging EfficientNet and Contrastive Learning for Accurate Global-scale Location Estimation. In *ICMR '21: International Conference on Multimedia Retrieval, Taipei, Taiwan, August 21-24, 2021.* 155–163.

[25] Liu Liu, Hongdong Li, and Yuchao Dai. 2019. Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization. In *Proc. of ICCV.* 2570–2579.

[26] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023. Dynamic LLM-Agent Network: An LLM-agent Collaboration Framework with Agent Team Optimization. *CoRR* (2023).

[27] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models. In *Proc. of NeurIPS.*

[28] Jiebo Luo, Dhiraj Joshi, Jie Yu, and Andrew C. Gallagher. 2011. Geotagging in multimedia and computer vision - a survey. *Multim. Tools Appl.* (2011), 187–211.

[29] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. 2018. Geolocation Estimation of Photos Using a Hierarchical Model and Scene Classification. In *Proc. of ECCV.* 575–592.

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proc. of NeurIPS.*

[31] Valerio Paolicelli, Gabriele Moreno Berton, Francesco Montagna, Carlo Masone, and Barbara Caputo. 2022. Adaptive-Attentive Geolocalization From Few Queries: A Hybrid Approach. *Frontiers Comput. Sci.* (2022), 841817.

[32] Guohao Peng, Yufeng Yue, Jun Zhang, Zhenyu Wu, Xiaoyu Tang, and Danwei Wang. 2021. Semantic Reinforced Attention Learning for Visual Place Recognition. In *Proc. of ICRA.* 13415–13422.

[33] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. 2022. Where in the world is this image? transformer-based geo-localization in the wild. In *Proc. of ECCV.* 196–215.

[34] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* (2019), 1655–1668.

[35] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are Emergent Abilities of Large Language Models a Mirage?. In *Proc. of NeurIPS.*

[36] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Proc. of NeurIPS.*

[37] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. 2018. CPlaNet: Enhancing Image Geolocalization by Combinatorial Partitioning of Maps. In *Proc. of ECCV.* 544–560.

[38] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. *CoRR* (2023).

[39] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Proc. of NeurIPS.*

[40] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. 2024. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Proc. of NeurIPS* (2024).

[41] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079* (2023).

[42] Xingyao Wang, Sha Li, and Heng Ji. 2022. Code4Struct: Code Generation for Few-Shot Structured Prediction from Natural Language. *CoRR* (2022).

[43] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. 2020. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proc. of CVPR.* 2626–2635.

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proc. of NeurIPS.*

[45] Jialiang Xu, Michael Moor, and Jure Leskovec. 2024. Reverse Image Retrieval Cues Parametric Memory in Multimodal LLMs. *arXiv preprint arXiv:2405.18740* (2024).

[46] Yang Yang, Hongchen Wei, Hengshu Zhu, Dianhai Yu, Hui Xiong, and Jian Yang. 2022. Exploiting cross-modal prediction and relation consistency for semisupervised image captioning. *IEEE Trans. on Cybernetics* (2022), 890–902.

[47] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proc. of ICLR.*

[48] Mubariz Zaffar, Sourav Garg, Michael Milford, Julian F. P. Kooij, David Flynn, Klaus D. McDonald-Maier, and Shoaib Ehsan. 2021. VPR-Bench: An Open-Source Visual Place Recognition Evaluation Framework with Quantifiable Viewpoint and Appearance Change. *Int. J. Comput. Vis.* (2021), 2136–2174.

[49] Lele Zhang, Fang Deng, Jie Chen, Yingcai Bi, Swee King Phang, Xudong Chen, and Ben M. Chen. 2018. Vision-Based Target Three-Dimensional Geolocation Using Unmanned Aerial Vehicles. *IEEE Trans. Ind. Electron.* (2018), 8052–8061.

[50] Chuanyang Zheng, Zhengying Liu, and Enze Xie et. al. 2023. Progressive-Hint Prompting Improves Reasoning in Language Models. *CoRR* (2023).

A Notations

We summarize all notations in this paper and list them in Table 6.

Table 6: Notations in this paper.

Notation	Description
X	The image to be recognized.
$Y\ (\hat{Y})$	The predicted (ground truth of) location.
$\mathcal{G}\ (\hat{\mathcal{G}})$	The predicted (ground truth of) LVLM social network.
$A\ (\hat{A})$	The predicted (ground truth of) adjacency matrix.
$Lst\ (\hat{Lst})$	The predicted (ground truth of) probability list.
$\mathcal{V}$	The set of LLM agents.
$\mathcal{E}$	The set of connections between LLM agents.
N	The number of agents.
K	The number of answer agent(s).
R	The number of review agent(s).
L	The number of agent discussion rounds.
Z	The maximum number of discussion rounds.
Θ	The learnable parameters in agent selection model.

B Implementation Details

As a supplement to the Implement Details in the main paper, for the agent social network learning model, we first deflate each image to be recognized to 512x512 pixels and then use the pre-trained VAE model[13] to embed the image. We define the embedding dimension to be 1024 and the hidden layer dimension of the network layer to be 1024. we use Adam as an optimizer for gradient descent with a learning rate of $1e^{-5}$. For each stage of the discussion, we use a uniform template for different agents and ask them to respond in a specified format. In addition, the model performance is the average of the last 100 epochs in a total training of 2500 epochs.

The detailed algorithm of smileGeo is illustrated in Algorithm 1. In the initialization stage, we initialize or load the parameters of the agent social network learning model (line 1). Next, we treat each LVLM agent as a node, establishing the LVLM agent collaboration social network and computing the adjacency relationships among LVLM agents as well as the probability that each agent is suited for responding to image X (line 2). Then, line 3 initializes the agent collaboration social network and line 4 computes the agent election probability. In Stage 1, line 5 involves electing appropriate answer agents based on the calculated probabilities. Subsequently, lines 6-10 detail the process through which each chosen answer agent formulates their response. Stage 2 begins by employing the random walk algorithm to assign review agents to each answer agent (lines 11-12). Lines 13-16 then describe how these review agents generate feedback based on the answers provided. In Stage 3, each answer agent consolidates feedback from their assigned review agents to finalize their response (lines 18-21). Line 22 concludes the final answer with up to Z rounds (we set $Z = 5$ in experiments) of intra-discussion among all answer agents only. The dynamic learning strategy module involves L-round (we set $L = 20$ in experiments) comparing the generated answers against the ground truth and updating the connections between the answer and review agents accordingly (lines 23-36). Finally, it concludes with the updating of the parameters of the dynamic learning model (line 37).

Algorithm 1 The smileGeo framework

Input: A set of pre-trained LLMs $\mathcal{V} = \{v_1, v_2, \cdots\}$, the input image X, and the ground truth $\hat{Y}$ (if has);
Output: The geospatial location Y.

Initialization Stage:
1: Initialize (Load) the parameter of the agent selection model: Θ
2: Calculate: $A \leftarrow f(X, \mathcal{V}|\Theta)$
3: Initialize the agent collaboration social network: $\mathcal{G}$
4: Calculate: $Lst \leftarrow f(X, \mathcal{G}|\Theta)$

Stage 1:
5: Elect K answer agents: $\mathcal{V}^1 = \{v_a^1, v_b^1, \cdots\} \leftarrow Lst$
6: **for** each answer agent v^1 **do**
7:     Obtain the location: $Y_{v^1}^1 \leftarrow \text{Ask}_{v^1}(X)$
8:     Get the confidence percentage: $C_{v^1}^1 \leftarrow \text{Ask}_{v^1}(X, Y_{v^1}^1)$
9:     Store the further explanation: $T_{v^1}^1 \leftarrow \text{Ask}_{v^1}(X, Y_{v^1}^1)$
10: **end for**

Stage 2:
11: **for** each selected answer agent v^1 **do**
12:     Select R review agents:
        $\mathcal{V}^2 = \{v_a^2, v_b^2, \cdots\} \leftarrow \text{RandomWalk}_{v^1}(\mathcal{G})$
13:     **for** each review agent v^2 **do**
14:         Obtain the comment $T_{v^2}^2 \leftarrow \text{Review}_{v^2}(X, Y_{v^1}^1, C_{v^1}^1)$
15:         Get the confidence percentage: $C_{v^2}^2 \leftarrow \text{Ask}_{v^2}(X, T_{v^2}^2)$
16:     **end for**
17: **end for**

Stage 3:
18: **for** each selected answer agent v^1 **do**
19:     Summary the final answer:
        $Y_{v^1}^3 \leftarrow \text{Summary}_{v^1}(Y_{v^1}^1, C_{v^1}^1, T_{v_1^2}^2, C_{v_1^2}^2, T_{v_2^2}^2, C_{v_2^2}^2, \cdots)$
20:     Get the final confidence percentage:
        $C_{v^1}^3 \leftarrow \text{Ask}_{v^1}(Y_{v^1}^1, C_{v^1}^1, T_{v_1^2}^2, C_{v_1^2}^2, T_{v_2^2}^2, C_{v_2^2}^2, \cdots)$
21: **end for**
22: Generate the final answer:
        $Y \leftarrow \text{Discussion}_Z(Y_{v_1^1}^3, C_{v_1^1}^3, Y_{v_2^1}^3, C_{v_2^1}^3, \cdots)$

The dynamic learning strategy module:
23: Initialize $Lst^{(0)}, \mathcal{G}^{(0)}$
24: **for** round l in total L rounds **do**
25:     **for** each selected answer agent v^1 **do**
26:         Obtain coordinates:
            $Coors \leftarrow \text{GeoEmb}(Y_{v^1}^3)$,
            $Coors_{\text{Truth}} \leftarrow \text{GeoEmb}(Y_{\text{Truth}})$
27:         **if** $\text{Dis}(Coors, Coors_{\text{Truth}}) \leq th$ **then**
28:             $A^{(l)} \leftarrow \text{Enhance}(e|e \text{ contains } v^1, e \in \mathcal{E})$
29:             Update $Lst^{(l)}[v^1] = 1$
30:         **else**
31:             $A^{(l)} \leftarrow \text{Weaken}(e|e \text{ contains } v^1, e \in \mathcal{E})$
32:             Update $Lst^{(l)}[v^1] = 0$
33:         **end if**
34:     **end for**
35: **end for**
36: $\hat{A} \approx A^{(L)}, \hat{Lst} \approx Lst^{(L)}$
37: Update: $\Theta \leftarrow Loss(\hat{Y}, Y, \hat{A}, A, \hat{Lst}, Lst)$

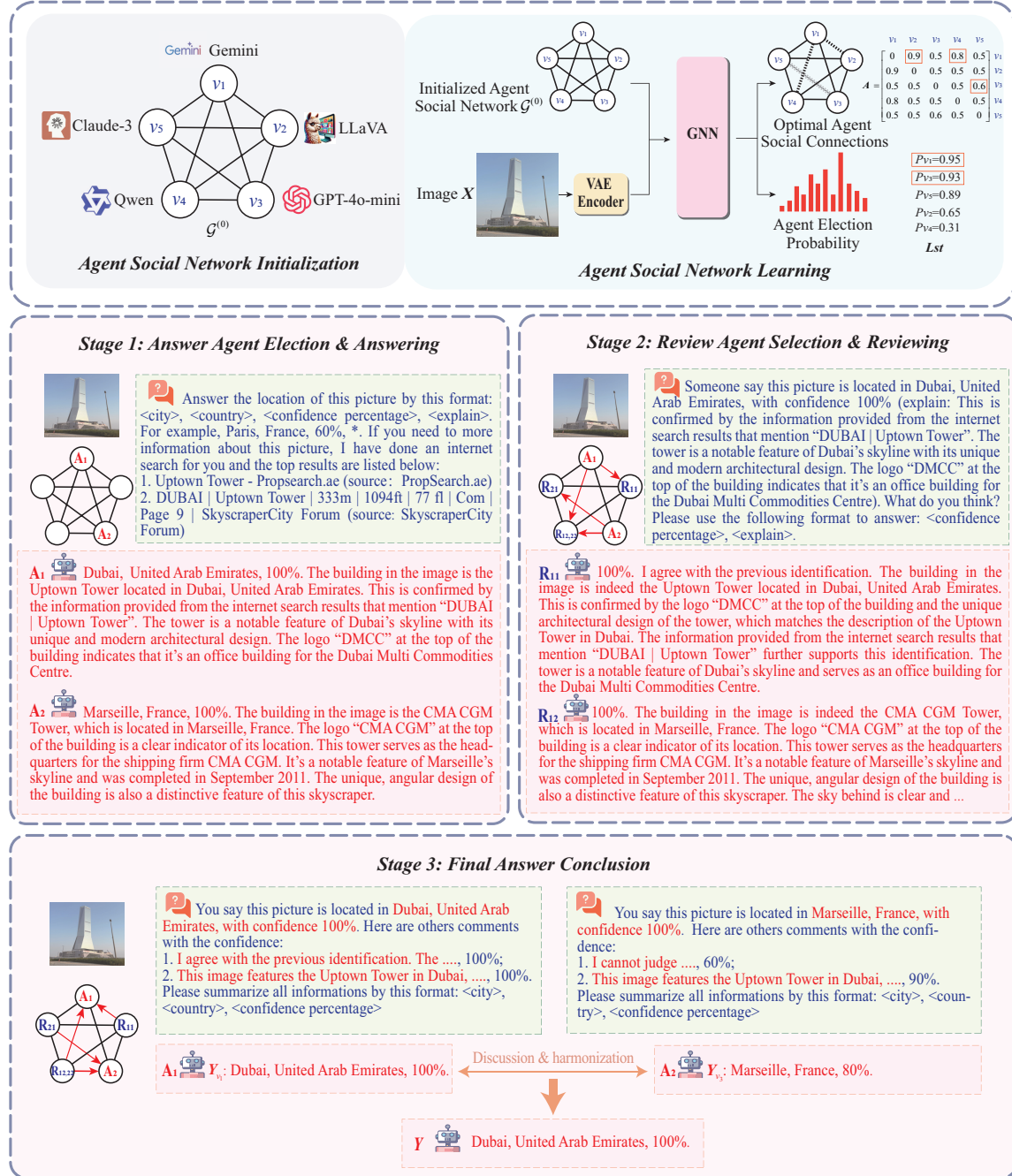[13]https://huggingface.co/stabilityai/sd-vae-ft-mse

Figure 8: A case study on the geo-localization process via a given image.

C Case Study

In Figure 8, we illustrate a case study of smileGeo. For this demonstration, we randomly select an image from the test dataset and employ five distinct LVLMs: LLaVA, GPT-4o-mini, Claude-3-opus, Gemini-1.5-pro, and Qwen2. The agent selection model selects two answer agents, as depicted in the top part of the figure. Then, stages 1-3 detail the process of generating the accurate location. Initially, only one answer agent provided the correct response. However, after several rounds of discussion, the agent that initially responded incorrectly revised its confidence level. During the final internal discussion, this agent aligned its response with the correct answer. This outcome validates the model efficiency, demonstrating its ability to integrate the knowledge and reasoning capabilities of different agents to enhance the overall performance.