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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Visual geo-localization demands in-depth knowledge and advanced reasoning skills to associate images with real-world geographic locations precisely. In general, traditional methods based on data-matching are hindered by the impracticality of storing adequate 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. Along this line, in this paper, we introduce a novel visual geo-localization framework called smileGeo that integrates the inherent knowledge of multiple LVLM agents via inter-agent communication to achieve effective geo-localization of images. Furthermore, our framework employs a dynamic learning strategy to optimize the communication patterns among agents, reducing unnecessary discussions among agents and improving the efficiency of the framework. To validate the effectiveness of the proposed framework, we construct GeoGlobe, a novel dataset for visual geo-localization tasks. Extensive testing on the dataset demonstrates that our approach significantly outperforms state-of-the-art methods. The source code is available at https://anonymous.4open.science/r/ViusalGeoLocalization-F8F5/ and the dataset will also be released after the paper is accepted.

1 Introduction

Visual geo-localization, referred to the task of estimating geographical identification for a given image, is vital in various fields such as human mobility analysis [1, 2, 3, 4, 5] and robotic navigation [6, 7, 8, 9, 10, 11]. In general, accurate visual geo-localization without the help of any localization equipment (*e.g.,* GPS sensors) is a complex task that requires abundant geospatial knowledge and strong reasoning capabilities. Traditional methods [12, 13, 14, 15] 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.

Recently, the success of Large Vision-Language Models (LVLMs) has enabled Visual Question Answering (VQA) to become a unified paradigm for multi-modal problems [16, 17], providing a novel solution for visual geo-localization without the need for external geo-tagged image records. However, the performance of a single LVLM on the geo-localization task is still limited by its inherent geospatial knowledge and reasoning capabilities. Along this line, in this paper, we introduce a novel multi-agent framework, named **s**war**m** **i**nte**l**lig**e**nce **Geo**-localization (**smileGeo**), which aims to adaptively integrate the inherent knowledge and reasoning capabilities of multiple LVLMs

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

to effectively and efficiently geo-localize images. Specifically, for a given image, the framework initially elects K suitable LVLM agents as answer agents for initial location analysis. Then, each answer agent chooses several review agents via an adaptive social network, which imitates the collaborative relationships between agents with a target on the visual geo-localization task, to discuss and share their knowledge for refining its location analysis. Finally, our framework conducts free discussion among all of the answer agents to reach a consensus. Besides, we also design a novel dynamic learning strategy to optimize the election mechanism along with the adaptive collaboration social network of agents. We hope that by the effectiveness of the election mechanism and the review mechanism, our framework can discover the mode of communication among agents, thereby enhancing geo-localization performance through multi-agent collaboration while minimizing unnecessary discussions. In summary, our contributions are demonstrated as follows:

- A novel swarm intelligence geo-localization framework, smileGeo, is proposed to adaptively integrate the inherent knowledge and reasoning capability of multiple LVLMs through discussion for visual geo-localization tasks.
- A dynamic learning strategy is introduced to discover the most appropriate discussion mode among LVLM agents for enhancing the effectiveness and efficiency of the framework.
- A new visual geo-localization dataset named GeoGlobe[1] is collected, containing a wide variety of images globally. The diversity and richness of GeoGlobe allow us to evaluate the performance of different models more accurately. Moreover, extensive experiments demonstrate our competitive performance compared to state-of-the-art methods.

The remainder of this paper is organized as follows: Section 2 discusses the related literature. In Section 3, the proposed framework is introduced. Section 4 provides the performance evaluation, and Section 5 concludes the paper.

2 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 [3, 18, 19, 20, 21, 22]. These systems utilize learned embeddings generated by a feature extraction backbone, which includes an aggregation or pooling mechanism [23, 24, 25, 26]. 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 [27, 28, 29, 30]. 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 [31, 32, 33], and a large number of studies have focused on developing robust architectures for collaborative LLM/LVLM agents [34, 35, 36, 37, 38]. These architectures enable each LLM/LVLM agent that endows with unique capabilities to engage in debates or discussions. For instance, [34] 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 [34] utilize a static architecture potentially limiting the performance and generalization of LLM/LVLM, others like [38] have implemented dynamic interaction architectures that adjust according to the query and incorporate user feedback.

[1] Because GeoGlobe is relatively large (about 32GB), we are unable to provide it as an attachment during the double-blind review stage. We will publish it once the paper is accepted.

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 [39, 40, 41], mathematical tools [40, 42, 43], and code interpreters [44, 45]. 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.

3 Methodology

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

3.1 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 edge set. 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 $\boldsymbol{Y}$ of a given image $\boldsymbol{X}$. 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 1, smileGeo contains the process of the review mechanism in agent discussions along with a dynamic learning strategy of agent social networks:

The review mechanism in agent discussions is a 3-stage anonymous collaboration approach to allow LVLM agents to reach a consensus via discussion. In the first stage, for a given image $\boldsymbol{X}$, our framework elects the most suitable K agents as answer agents by agent election probability $\boldsymbol{Lst}$. In the second stage, these answer agents respectively select R review agents by the adaptive collaboration social network $\boldsymbol{A}$ to refine their answer via discussion. Finally, our 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 $\boldsymbol{X}$, 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 $\boldsymbol{Y}$.

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

3.2 Review Mechanism in Agent Discussions

LLM/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 LLM/LVLM agents. Thus, to improve the geo-localization capability of LVLMs, we propose a cooperation framework to effectively integrate the diverse knowledge 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

Initially, we select K agents with the highest agent election probabilities $\boldsymbol{Lst}$ 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.

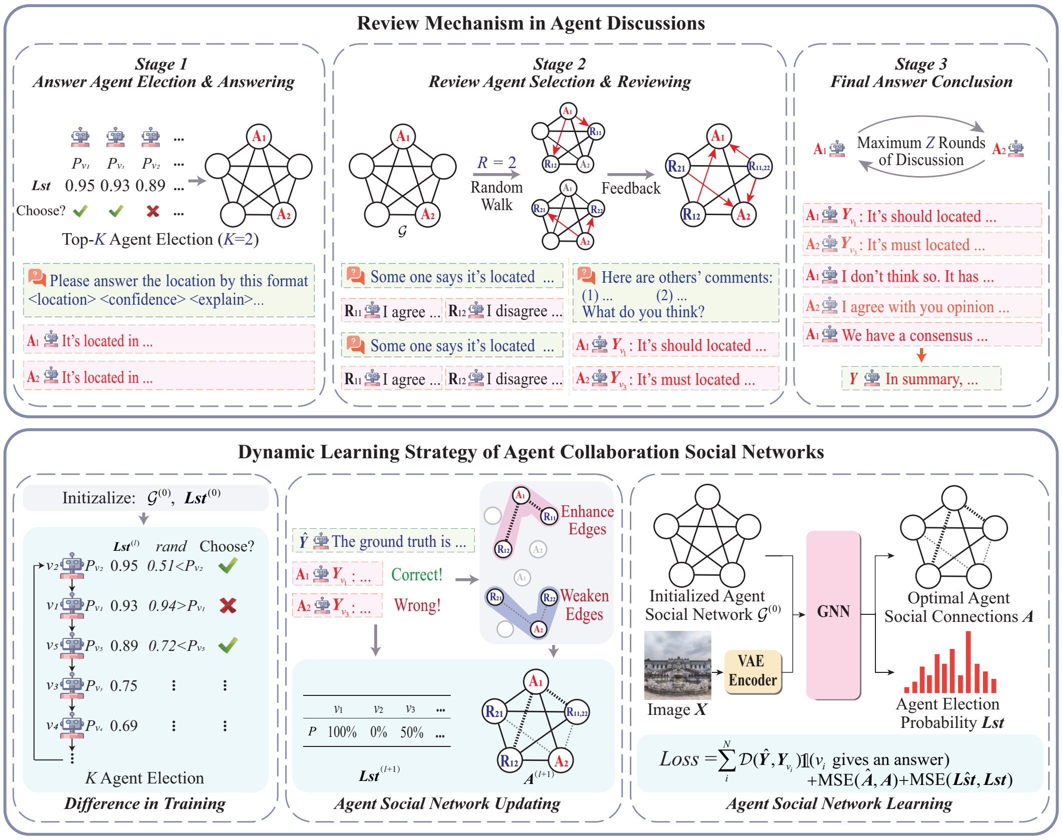


Figure 1: The framework overview of smileGeo. It contains the process of review mechanism in agent discussions along with a dynamic learning strategy of agent collaboration social networks. The first part deploys a review mechanism for LVLMs to discuss and share their knowledge anonymously, which could 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.

After the answer agents are elected, we send the image $\boldsymbol{X}$ 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.

Stage 2: Review Agent Selection & Reviewing

In this stage, for each answer agent, we choose R 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:

$$p(v_i, v_j) = \begin{cases} \frac{\boldsymbol{A}_{ij}}{\sum_{k \in \mathcal{N}(v_i)} \boldsymbol{A}_{ik}}, & \text{if } e_{ij} \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

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 explain.

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, 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.

3.3 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 $\hat{\boldsymbol{Lst}}$ 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 $\boldsymbol{Lst}$ and $\mathcal{G}$, by such estimated ground truth.

To estimate the optimal $\hat{\boldsymbol{Lst}}$ and $\hat{\boldsymbol{A}}$ for agents to geo-localize image $\boldsymbol{X}$, 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 $\boldsymbol{Lst}^{(0)} = [0.5, 0.5, \cdots]$, 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. $\boldsymbol{Lst}^{(l)}$ and $\mathcal{G}^{(l)}$ is updated at the end of each round $l \in L$ by comparing the answers $\boldsymbol{Y}_{v_i}^{(l)}$ from each answer agent with the ground truth $\hat{\boldsymbol{Y}}$.

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

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

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 1.

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:

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

where $\boldsymbol{A}_{ij}^{(l-1)}(\boldsymbol{X})$ is the weight of the connection between answer agent v_i and review agent v_j at round $l-1$ when geo-locating image $\boldsymbol{X}$, $\boldsymbol{A}_{ij}^{(0)}(\boldsymbol{X}) = 1, i \neq j, \boldsymbol{A}_{ii}^{(0)}(\boldsymbol{X}) = 0, i, j \in [N]$, 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 $\boldsymbol{Lst} = h(\boldsymbol{X}, \mathcal{G}|\Theta)$ and the optimal agent collaboration connections $\boldsymbol{A} = f(\boldsymbol{X}, \mathcal{V}|\Theta)$:

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

where $\boldsymbol{W}, \boldsymbol{Emb} \in \Theta$ are two learnable parameters, $\boldsymbol{Emb} := [\boldsymbol{Emb}_{v_1}, \boldsymbol{Emb}_{v_2}, \cdots]^\top$ is the agent embedding and $\boldsymbol{W}$ is the weight matrix, $\sigma(\cdot)$ is the LeakyReLU function, $\sigma'(\cdot)$ is the Sigmoid function, $\mathrm{VAE}_{\mathrm{Enc}}(\cdot)$ 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 $\boldsymbol{Fea}$. Our learning target can be formalized as:

$$\arg\min_{\Theta} \sum_{i}^{N} \mathcal{D}(\hat{\boldsymbol{Y}}, \boldsymbol{Y}_{v_i}) \mathbb{1}(v_i \text{ gives an answer}) + \mathrm{MSE}(\hat{\boldsymbol{Lst}}, \boldsymbol{Lst}) + \mathrm{MSE}(\hat{\boldsymbol{A}}, \boldsymbol{A}), \tag{5}$$

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, $\boldsymbol{Y}_{v_i} := \boldsymbol{Y}_{v_i}^{(L)} = g_{v_i}(\boldsymbol{X}, \boldsymbol{Y}_{v_j}^{(L-1)})$, $g_{v_i}(\cdot)$ represent the LVLM agent v_i with fixed parameters and $\boldsymbol{Y}_{v_i}^{(0)} = g_{v_i}(\boldsymbol{X})$ is the answer that LVLM agent v_i generates at the initial stage of discussion.

4 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 smileGeo outperform state-of-the-art methods in open-ended geo-localization tasks?
- **RQ2**: Are LVLM agents with diverse knowledge and reasoning abilities more suitable for building a collaboration social network of agents?
- **RQ3**: How does the setting of hyperparameters affect the performance of smileGeo?

4.1 Experiment Setup

Datasets. In this paper, we newly construct a geo-localization dataset named GeoGlobe. 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. More details can be found in Appendix B.

Implemention Details. We select both open-source and close-source LVLMs with different scales trained by different datasets as agents in the proposed framework. As for the open-source LVLMs, we utilize several open-source fine-tuned LVLMs: Infi-MM[2], Qwen-VL [3], vip–llava–7b&13b[4], llava–

[2]https://huggingface.co/Infi-MM/infimm-zephyr

[3]https://huggingface.co/Qwen/Qwen-VL

[4]https://huggingface.co/llava-hf/vip-llava-xxx

1.5–7b–base&mistral&vicuna[5], llava–1.6–7b&13b&34b–mistral&vicuna[6], CogVLM[7]. As for the closed-source LVLMs, we chose the models provided by three of the most famous companies in the world: Claude–3–opus[8], GPT–4V[9], and Gemini–1.5–pro [10]. Besides, 99% of images (about 290,000 samples) from the original dataset are randomly chosen as training samples. For the open-world geolocation problem, we construct the test dataset using approximately 4,000 samples, of which nearly 66.67% samples reflected different locations not present in the training dataset. More details about the deployment of smileGeo and the related parameter settings can be found in Appendix C.

Baselines. In this work, we compare the proposed framework with three kinds of baselines: single LVLMs, LLM/LVLM-based multi-agent frameworks, and image retrieval approaches. Firstly, we use each LVLM alone as an agent directly for the geo-localization task and compute the performance of these single LVLMs under the same dataset. In addition, we experiment with multi-agent collaborative frameworks, including LLM-Blender [34], PHP [35], Reflexion [36], LLM Debate [37], and DyLAN [38]. Finally, several state-of-the-art image retrieval approaches, including NetVLAD [3], GeM [26], and CosPlace [46], are also used to be part of the baselines. We set the training dataset as the geo-tagged image database of each image retrieval system and use images in the test dataset for the retrieval system to generate answers.

Evaluation Metrics. We use *Accuracy* (Acc) to evaluate the performance: $Accuracy = \frac{N_{\text{correct}}}{N_{\text{total}}}$, 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.

In this paper, we first geo-encode the answers with the ground truth, *i.e.,* we transform the addresses described through natural language into latitude-longitude coordinates. Then, we calculate the distance between the two coordinates. When the distance between the two coordinates is less than $th = 50km$ (city-level), we consider the answer of the framework to be correct.

4.2 Performance Comparison

We divide the baseline comparison experiment into three parts: i) comparison with single LVLMs, ii) comparison with LLM/LVLM-based agent frameworks, and iii) comparison with image retrieval systems.

Table 1: Results of different single LVLM baselines.

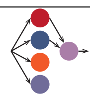
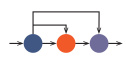
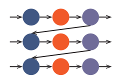
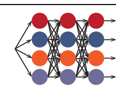
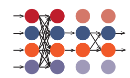
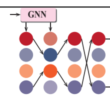
	Without Web Searching			**With Web Searching**		
	Natural	**ManMade**	**Overall**	**Natural**	**ManMade**	**Overall**
Infi-MM	19.2547	21.4133	20.9883	0.9938	0.3351	0.4648
Qwen-VL	42.4845	37.4657	38.4540	4.9689	11.2093	9.9804
vip-llava-13b	20.6211	15.4127	16.4384	8.323	4.3558	5.137
vip-llava-7b	21.9876	18.4892	19.1781	31.9255	56.5032	51.6634
llava-1.5-7b	17.3913	16.3265	16.5362	27.205	47.2129	43.273
llava-1.6-7b-mistral	0.3727	0.0914	0.1468	0.8696	2.1627	1.908
llava-1.6-7b-vicuna	2.2360	2.0713	2.1037	6.9565	15.8696	14.1145
llava-1.6-13b	10.4348	8.8943	9.1977	12.1739	28.2668	25.0978
llava-1.6-34b	10.3106	9.1379	9.3689	52.795	77.1855	72.3826
CogVLM	7.7019	7.5845	7.6076	6.8323	10.3564	9.6624
claude-3-opus	22.06	37.38	16.5468	33.0435	40.7125	39.2027
GPT-4V	27.5776	35.3443	33.8145	61.9876	87.6028	82.5587
Gemini-1.5-pro	55.6522	60.3107	59.3933	62.2360	82.8206	78.7671
smileGeo	**58.6111**	**64.3968**	**63.2730**	**78.0448**	**87.0069**	**85.2630**

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

[5]https://huggingface.co/llava-hf/llava-1.5-xxx
[6]https://huggingface.co/liuhaotian/llava-v1.6-xxx
[7]https://github.com/THUDM/CogVLM
[8]https://anthropic.com/
[9]https://openai.com/
[10]https://gemini.google.com/

Firstly, the performance of all single LVLM baselines is shown in Table 1, in terms of the metric Acc. The data in Table 1 indicate that open-source LVLMs with diverse knowledge and reasoning capabilities exhibit significant variations, particularly in geo-localization tasks. This may be due to the difference in the overlap between the pre-training datasets used by different LVLMs and the dataset we constructed. Therefore, in addition to querying the LVLM locations about images, we also incorporated real-time image search results from Google to provide the model with more comprehensive information. These results from Internet retrievals are incorporated into the chain-of-thoughts (CoT) [47] of LVLMs as external knowledge. At this time, models with larger parameters, such as llava–1.6–34b, demonstrate superior reasoning abilities compared to smaller models (7b or 13b). In addition, closed-source large models also show more consistent performance than their open-source counterparts and are more adept at analyzing and utilizing external knowledge for accurate inferences. Compared to all single LVLMs, our proposed LVLM agent framework surpasses all single LVLM baselines in accuracy. This improvement confirms the effectiveness of different LVLMs collaborating by engaging in discussions and analyzing various types of images, thus producing more precise results.

Table 2: Results of different agent frameworks without web searching.

Framework	LLM-Blender	PHP	Reflexion	LLM Debate	DyLAN	**smileGeo**
Sturcture						
Acc ↑	55.7802%	60.9809%	62.3412%	57.0119%	62.8187%	**63.2730%**
Tks ↓	23,662	154,520	109,524	260,756	159,320	**18,876**

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

Secondly, the comparative results across various LLM/LVLM agent frameworks are presented in Table 2. It is evident that the majority of LLM/LVLM agent frameworks surpass individual LVLMs in terms of geo-localization accuracy. 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 18,876, 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.

Finally, Table 3 presents the comparison between the proposed framework and existing image retrieval systems. Our framework, smileGeo, consistently outperforms all other retrieval-based approaches. This superior performance can be attributed to the fact that other image retrieval methods rely on a rich geo-tagged image database. In our test dataset, however, two-thirds of the images are new and localized in completely different areas from those in the training dataset. This highlights the shortages of conventional database-based retrieval systems due to the limitations of the geo-tagged image databases and demonstrates the effectiveness of our proposed framework in solving open-world geo-localization tasks.

Table 3: Comparison with image retrieval systems.

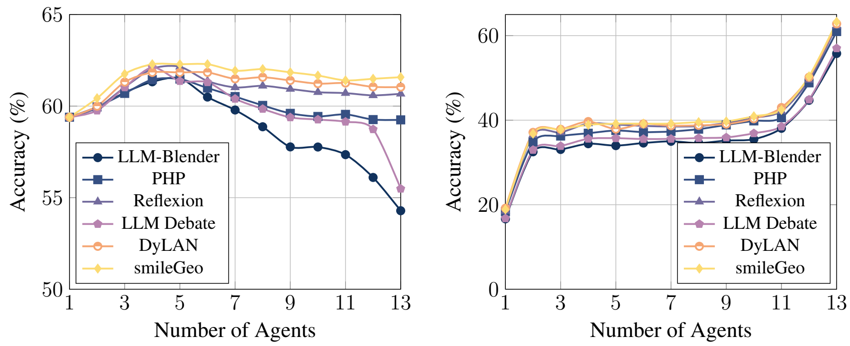
	Natural	**ManMade**	**Overall**
NetVLAD	26.5134	28.9955	28.6047
GeM	23.1022	25.4175	25.0749
CosPlace	28.1688	30.2782	29.8701
smileGeo	**58.6111**	**64.3968**	**63.2730**

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

4.3 Ablation Study

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 2.



(a) Calling the same LVLM API

(b) Different LVLM backbones

Figure 2: Results of model performance in relation to the number of agents.

As illustrated in Figure 2(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 models with the same 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 2(b) reveals that the accuracy of the framework improves with the incorporation of more LVLM backbones, indicating that the diversity of LVLMs 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 3. 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 balance both the efficiency and the accuracy of smileGeo, for the experiments presented in this paper, we set both K and R equal to 2.

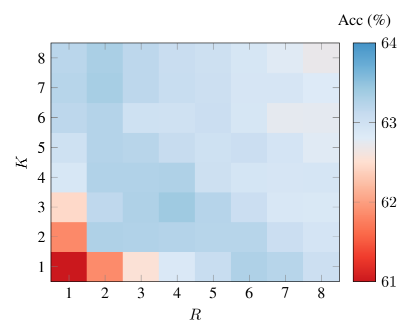


Figure 3: Results under different K and R.

5 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 and will open-source it. Overall, smileGeo demonstrates significant improvements in geo-localization tasks, achieving superior performance with lower computational demands compared to contemporary state-of-the-art LLM/LVLM agent frameworks.

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 complex scenarios, such as high-precision global positioning and navigation for robots, laying the cornerstone for exploring LVLM agent collaboration to handle different complex open-world tasks efficiently.

References

[1] B. Huang and K. M. Carley, "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*, F. Spezzano, W. Chen, and X. Xiao, Eds. ACM, 2019, pp. 365–373. [Online]. Available: https://doi.org/10.1145/3341161.3342870

[2] J. Luo, D. Joshi, J. Yu, and A. C. Gallagher, "Geotagging in multimedia and computer vision - a survey," *Multim. Tools Appl.*, vol. 51, no. 1, pp. 187–211, 2011. [Online]. Available: https://doi.org/10.1007/s11042-010-0623-y

[3] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: CNN architecture for weakly supervised place recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1437–1451, 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2711011

[4] M. Zaffar, S. Garg, M. Milford, J. F. P. Kooij, D. Flynn, K. D. McDonald-Maier, and S. Ehsan, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2136–2174, 2021. [Online]. Available: https://doi.org/10.1007/s11263-021-01469-5

[5] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 2, pp. 257–271, 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2017.2667665

[6] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. D. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *2017 IEEE International Conference on Robotics and Automation, ICRA 2017, Singapore, Singapore, May 29 - June 3, 2017*. IEEE, 2017, pp. 3223–3230. [Online]. Available: https://doi.org/10.1109/ICRA.2017.7989366

[7] Z. Chen, L. Liu, I. Sa, Z. Ge, and M. Chli, "Learning context flexible attention model for long-term visual place recognition," *IEEE Robotics Autom. Lett.*, vol. 3, no. 4, pp. 4015–4022, 2018. [Online]. Available: https://doi.org/10.1109/LRA.2018.2859916

[8] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*. IEEE, 2017, pp. 9–16. [Online]. Available: https://doi.org/10.1109/IROS.2017.8202131

[9] S. Garg, N. Sünderhauf, and M. Milford, "Semantic-geometric visual place recognition: a new perspective for reconciling opposing views," *Int. J. Robotics Res.*, vol. 41, no. 6, pp. 573–598, 2022. [Online]. Available: https://doi.org/10.1177/0278364919839761

[10] S. Hausler, A. Jacobson, and M. Milford, "Multi-process fusion: Visual place recognition using multiple image processing methods," *IEEE Robotics Autom. Lett.*, vol. 4, no. 2, pp. 1924–1931, 2019. [Online]. Available: https://doi.org/10.1109/LRA.2019.2898427

[11] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. D. McDonald-Maier, "A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes," *IEEE Trans. Robotics*, vol. 36, no. 2, pp. 561–569, 2020. [Online]. Available: https://doi.org/10.1109/TRO.2019.2956352

[12] M. M. ElQadi, M. Lesiv, A. G. Dyer, and A. Dorin, "Computer vision-enhanced selection of geo-tagged photos on social network sites for land cover classification," *Environ. Model. Softw.*, vol. 128, p. 104696, 2020. [Online]. Available: https://doi.org/10.1016/j.envsoft.2020.104696

[13] M. Campbell and M. Wheeler, "A vision based geolocation tracking system for uav's," in *AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2006, p. 6246.

[14] F. Deng, L. Zhang, F. Gao, H. Qiu, X. Gao, and J. Chen, "Long-range binocular vision target geolocation using handheld electronic devices in outdoor environment," *IEEE Trans. Image Process.*, vol. 29, pp. 5531–5541, 2020. [Online]. Available: https://doi.org/10.1109/TIP.2020.2984898

[15] L. Zhang, F. Deng, J. Chen, Y. Bi, S. K. Phang, X. Chen, and B. M. Chen, "Vision-based target three-dimensional geolocation using unmanned aerial vehicles," *IEEE Trans. Ind. Electron.*, vol. 65, no. 10, pp. 8052–8061, 2018. [Online]. Available: https://doi.org/10.1109/TIE.2018.2807401

[16] X. Feng, Z.-Y. Chen, Y. Qin, Y. Lin, X. Chen, Z. Liu, and J.-R. Wen, "Large language model-based human-agent collaboration for complex task solving," *arXiv preprint arXiv:2402.12914*, 2024.

[17] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song *et al.*, "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.

[18] V. Paolicelli, G. M. Berton, F. Montagna, C. Masone, and B. Caputo, "Adaptive-attentive geolocalization from few queries: A hybrid approach," *Frontiers Comput. Sci.*, vol. 4, p. 841817, 2022. [Online]. Available: https://doi.org/10.3389/fcomp.2022.841817

[19] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li, "Self-supervising fine-grained region similarities for large-scale image localization," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12349. Springer, 2020, pp. 369–386. [Online]. Available: https://doi.org/10.1007/978-3-030-58548-8_22

[20] H. Jin Kim, E. Dunn, and J.-M. Frahm, "Learned contextual feature reweighting for image geo-localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2136–2145.

[21] L. Liu, H. Li, and Y. Dai, "Stochastic attraction-repulsion embedding for large scale image localization," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 2570–2579. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00266

[22] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2626–2635.

[23] G. Peng, Y. Yue, J. Zhang, Z. Wu, X. Tang, and D. Wang, "Semantic reinforced attention learning for visual place recognition," in *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*. IEEE, 2021, pp. 13 415–13 422. [Online]. Available: https://doi.org/10.1109/ICRA48506.2021.9561812

[24] S. Ibrahimi, N. van Noord, T. Alpherts, and M. Worring, "Inside out visual place recognition," in *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*. BMVA Press, 2021, p. 362. [Online]. Available: https://www.bmvc2021-virtualconference.com/assets/papers/0467.pdf

[25] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 14 141–14 152. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Hausler_Patch-NetVLAD_Multi-Scale_Fusion_of_Locally-Global_Descriptors_for_Place_Recognition_CVPR_2021_paper.html

[26] F. Radenovic, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, 2019. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2846566

[27] M. Izbicki, E. E. Papalexakis, and V. J. Tsotras, "Exploiting the earth's spherical geometry to geolocate images," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019, Würzburg, Germany, September 16-20, 2019, Proceedings, Part II*, ser. Lecture Notes in Computer Science, U. Brefeld, É. Fromont, A. Hotho, A. J. Knobbe, M. H. Maathuis, and C. Robardet, Eds., vol. 11907. Springer, 2019, pp. 3–19. [Online]. Available: https://doi.org/10.1007/978-3-030-46147-8_1

[28] G. Kordopatis-Zilos, P. Galopoulos, S. Papadopoulos, and I. Kompatsiaris, "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*, W. Cheng, M. S. Kankanhalli, M. Wang, W. Chu, J. Liu, and M. Worring, Eds. ACM, 2021, pp. 155–163. [Online]. Available: https://doi.org/10.1145/3460426.3463644

[29] E. Müller-Budack, K. Pustu-Iren, and R. 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*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11216. Springer, 2018, pp. 575–592. [Online]. Available: https://doi.org/10.1007/978-3-030-01258-8_35

[30] P. H. Seo, T. Weyand, J. Sim, and B. Han, "Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11214. Springer, 2018, pp. 544–560. [Online]. Available: https://doi.org/10.1007/978-3-030-01249-6_33

[31] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[32] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. M. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of artificial general intelligence: Early experiments with GPT-4," *CoRR*, vol. abs/2303.12712, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.12712

[33] R. Schaeffer, B. Miranda, and S. Koyejo, "Are emergent abilities of large language models a mirage?" in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/adc98a266f45005c403b8311ca7e8bd7-Abstract-Conference.html

[34] D. Jiang, X. Ren, and B. Y. Lin, "Llm-blender: Ensembling large language models with pairwise ranking and generative fusion," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 14 165–14 178. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.792

[35] C. Zheng, Z. Liu, E. Xie, Z. Li, and Y. Li, "Progressive-hint prompting improves reasoning in large language models," *CoRR*, vol. abs/2304.09797, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.09797

[36] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: language agents with verbal reinforcement learning," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html

[37] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," *CoRR*, vol. abs/2305.14325, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.14325

[38] Z. Liu, Y. Zhang, P. Li, Y. Liu, and D. Yang, “Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization,” *CoRR*, vol. abs/2310.02170, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.02170

[39] W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W. Yih, “REPLUG: retrieval-augmented black-box language models,” *CoRR*, vol. abs/2301.12652, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.12652

[40] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, “React: Synergizing reasoning and acting in language models,” in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/pdf?id=WE_vluYUL-X

[41] G. Izacard, P. S. H. Lewis, M. Lomeli, L. Hosseini, F. Petroni, T. Schick, J. Dwivedi-Yu, A. Joulin, S. Riedel, and E. Grave, “Atlas: Few-shot learning with retrieval augmented language models,” *J. Mach. Learn. Res.*, vol. 24, pp. 251:1–251:43, 2023. [Online]. Available: http://jmlr.org/papers/v24/23-0037.html

[42] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, “Toolformer: Language models can teach themselves to use tools,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html

[43] P. Lu, B. Peng, H. Cheng, M. Galley, K. Chang, Y. N. Wu, S. Zhu, and J. Gao, “Chameleon: Plug-and-play compositional reasoning with large language models,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/871ed095b734818cfba48db6aeb25a62-Abstract-Conference.html

[44] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig, “PAL: program-aided language models,” in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 10 764–10 799. [Online]. Available: https://proceedings.mlr.press/v202/gao23f.html

[45] X. Wang, S. Li, and H. Ji, “Code4struct: Code generation for few-shot structured prediction from natural language,” *CoRR*, vol. abs/2210.12810, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2210.12810

[46] G. M. Berton, C. Masone, and B. Caputo, “Rethinking visual geo-localization for large-scale applications,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 4868–4878. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.00483

[47] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

A Notations

We summarize all notations in this paper and list them in Table 4.

Table 4: Notations in this paper.

Notation	Description
$\boldsymbol{X}$	The image to be recognized.
$\boldsymbol{Y}$ ($\hat{\boldsymbol{Y}}$)	The predicted (ground truth of) geospatial location in the natural language form.
$\mathcal{G}$ ($\hat{\mathcal{G}}$)	The predicted (ground truth of) LVLM-based agent collaboration social network.
$\boldsymbol{A}$ ($\hat{\boldsymbol{A}}$)	The predicted (ground truth of) adjacency matrix of the agent social network.
$\boldsymbol{Lst}$ ($\hat{\boldsymbol{Lst}}$)	The predicted (ground truth of) scalar of agent election probability.
$\mathcal{V}$	The set of LLM agents.
$\mathcal{E}$	The set of connections between LLM agents.
N	The number of agents.
K	The number of agents to be elected as answer agent(s).
R	The number of agents to be selected as review agent(s).
L	The number of agent discussion rounds.
Z	The maximum number of rounds in which answer agents harmonize opinions.
Θ	The learnable parameters of the agent social network learning model.

B Dataset Details

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 as well as 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 5 and Figure 4 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.

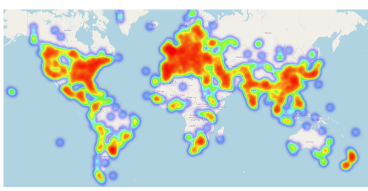


Figure 4: The data distribution around the world.

Table 5: Statistics of the dataset GeoGlobe.

	Images	Cities	Countries	Attractions
Man-made	253,118	2,313	143	10,492
Natural	40,087	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 training samples as $\mathcal{Z}_{\text{train}}$, and the 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 $consistency \approx 33.2957\%$.

C Implementation Details

In all experiments, we employ a variety of LVLMs, encompassing both open-source and closed-source models, to be agents in the proposed framework. Unless specified otherwise, zero-shot prompting is applied. Each open-source LVLM is deployed on a dedicated A800 (80G) GPU server with 200GB memory. As for each closed-source LVLM, we cost amounting to billions of tokens by calling APIs as specified by the official website. 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

Algorithm 1 The smileGeo framework

Input: A set of pre-trained LLMs $\mathcal{V} = \{v_1, v_2, \cdots\}$, the input image $\boldsymbol{X}$, and the ground truth $\hat{\boldsymbol{Y}}$ (if has);
Output: The geospatial location $\boldsymbol{Y}$.
Initialization Stage:
1: Initialize (Load) the parameter of the agent selection model: Θ
2: Calculate: $\boldsymbol{A} \leftarrow f(X, \mathcal{V}|\Theta)$
3: Initialize the agent collaboration social network: $\mathcal{G}$
4: Calculate: $\boldsymbol{Lst} \leftarrow f(X, \mathcal{G}|\Theta)$
Stage 1:
5: Elect answer agents: $\mathcal{V}^1 = \{v_a^1, v_b^1, \cdots\} \leftarrow \boldsymbol{Lst}$, where $|\mathcal{V}^1| = K$
6: **for** each answer agent v^1 **do**
7: Obtain the location: $\boldsymbol{Y}^1_{v^1} \leftarrow \mathrm{Ask}_{v^1}(\boldsymbol{X})$
8: Get the confidence percentage: $C^1_{v^1} \leftarrow \mathrm{Ask}_{v^1}(\boldsymbol{X}, \boldsymbol{Y}^1_{v^1})$
9: Store the further explanation: $T^1_{v^1} \leftarrow \mathrm{Ask}_{v^1}(\boldsymbol{X}, \boldsymbol{Y}^1_{v^1})$
10: **end for**
Stage 2:
11: **for** each selected answer agent v^1 **do**
12: Select the review agents: $\mathcal{V}^2 = \{v_a^2, v_b^2, \cdots\} \leftarrow \mathrm{RandomWalk}_{v^1}(\mathcal{G})$, where $|\mathcal{V}^2| = R$
13: **for** each review agent v^2 **do**
14: Obtain the comment $T^2_{v^2} \leftarrow \mathrm{Review}_{v^2}(\boldsymbol{X}, \boldsymbol{Y}^1_{v^1}, C^1_{v^1})$
15: Get the confidence percentage: $C^2_{v^2} \leftarrow \mathrm{Ask}_{v^2}(\boldsymbol{X}, T^2_{v^2})$
16: **end for**
17: **end for**
Stage 3:
18: **for** each selected answer agent v^1 **do**
19: Summary the final answer: $\boldsymbol{Y}^3_{v^1} \leftarrow \mathrm{Summary}_{v^1}(\boldsymbol{Y}^1_{v^1}, C^1_{v^1}, T^2_{v^2_1}, C^2_{v^2_1}, T^2_{v^2_2}, , C^2_{v^2_2}, \cdots)$
20: Get the final confidence percentage: $C^3_{v^1} \leftarrow \mathrm{Ask}_{v^1}(\boldsymbol{Y}^1_{v^1}, C^1_{v^1}, T^2_{v^2_1}, C^2_{v^2_1}, T^2_{v^2_2}, , C^2_{v^2_2}, \cdots)$
21: **end for**
22: Generate the final answer: $\boldsymbol{Y} \leftarrow \mathrm{Discussion}_Z(\boldsymbol{Y}^3_{v^1_1}, C^3_{v^1_2}, \boldsymbol{Y}^3_{v^1_2}, C^3_{v^1_2}, \cdots)$
The dynamic learning strategy module:
23: Initialize $\boldsymbol{Lst}^{(0)}, \mathcal{G}^{(0)}$
24: **for** round l in total L rounds **do**
25: **for** each selected answer agent v^1 **do**
26: Obtain coordinates: $Coors \leftarrow \mathrm{GeoEmb}(\boldsymbol{Y}^3_{v^1})$, $Coors_{\text{Truth}} \leftarrow \mathrm{GeoEmb}(\boldsymbol{Y}_{\text{Truth}})$
27: **if** $\mathrm{Dis}(Coors, Coors_{\text{Truth}}) \leq th$ **then**
28: $\boldsymbol{A}^{(l)} \leftarrow \mathrm{Enhance}(e|e \text{ contains } v^1, e \in \mathcal{E})$
29: Update $\boldsymbol{Lst}^{(l)}[v^1] = 1$
30: **else**
31: $\boldsymbol{A}^{(l)} \leftarrow \mathrm{Weaken}(e|e \text{ contains } v^1, e \in \mathcal{E})$
32: Update $\boldsymbol{Lst}^{(l)}[v^1] = 0$
33: **end if**
34: **end for**
35: **end for**
36: $\hat{\boldsymbol{A}} \approx \boldsymbol{A}^{(L)}$, $\hat{\boldsymbol{Lst}} \approx \boldsymbol{Lst}^{(L)}$
37: Update: $\Theta \leftarrow Loss(\hat{\boldsymbol{Y}}, \boldsymbol{Y}, \hat{\boldsymbol{A}}, \boldsymbol{A}, \hat{\boldsymbol{Lst}}, \boldsymbol{Lst})$

tokens that each agent supports. Besides, noting that images are token consuming, we only keep the freshest response for agent discussions.

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, as delineated in 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 $\boldsymbol{X}$, as shown in line 2. Then, line 3 initializes the agent

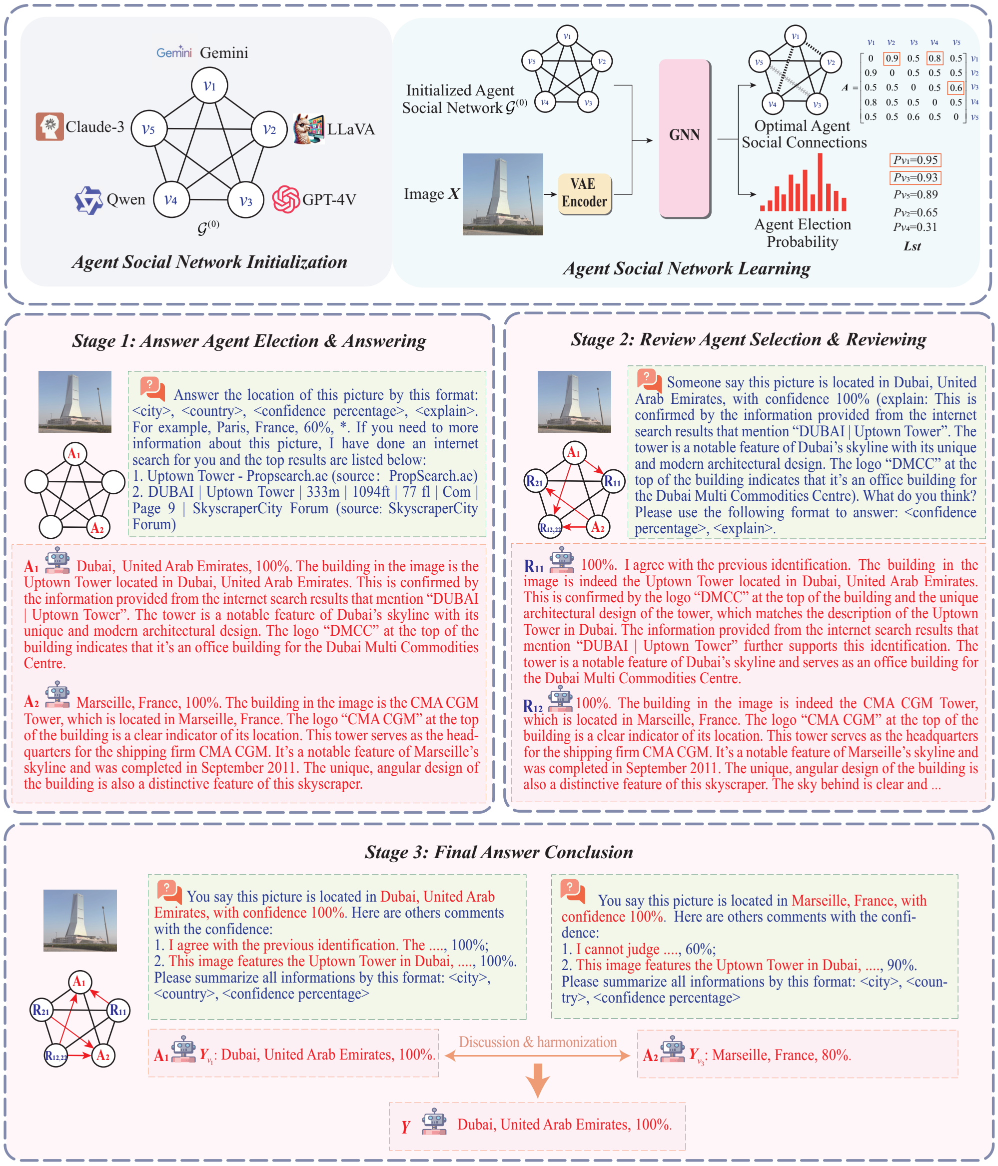
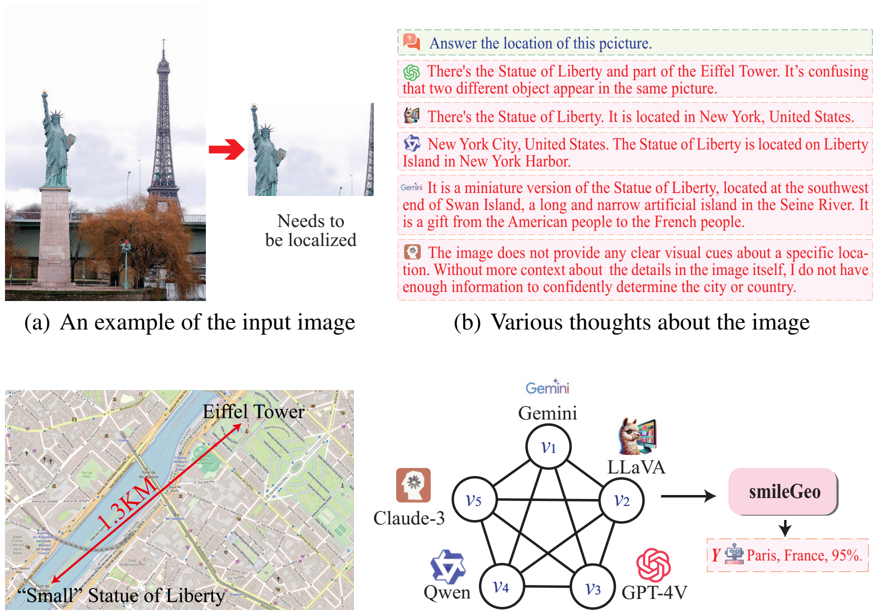


Figure 5: A case study on the geo-localization process via a given image.

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, as depicted in 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, as illustrated in lines 18-21. Line 22 concludes the final answer with up to Z rounds (we set $Z = 10$ 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, as shown in lines 23-36. In line 37, the process concludes with the updating of the learning parameters of the dynamic agent social network learning model.

Here, 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[11] to compress the image again (compression

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



(a) An example of the input image (b) Various thoughts about the image

(c) Actual locations of two landmarks (d) The final answer of smileGeo

Figure 6: A case study illustrating the reasoning capabilities of smileGeo.

ratio 1:8) and extract its representations. We define the embedding dimension of the nodes 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 LVLM agent discussion, we use a uniform template to ask questions to different LVLM agents and ask them to make a response in the specified format. In addition, the performance of our proposed framework is the average of the last 100 epochs in a total training of 2500 epochs.

D Additional Experiments

D.1 Case Study

Case 1: In Figure 5, we illustrate the application of smileGeo in a visual geo-localization task. For this demonstration, we randomly select an image from the test dataset and employ five distinct LVLMs: LLaVA, GPT-4V, Claude-3, Gemini, and Qwen. The agent selection model selects two answer agents, as depicted in the top part of the figure. Subsequently, stages 1 through 3 detail the process of generating the accurate geo-location. Initially, only one answer agent provided the correct response. However, after several rounds of discussion, the agent that initially responded incorrectly revised the confidence level of its answer. During the final internal discussion, this agent aligned its response with the correct answer. This outcome validates the efficacy of our proposed framework, demonstrating its ability to integrate the knowledge and reasoning capabilities of different agents to enhance the overall performance of the proposed LVLM agent framework.

Case 2: This case study illustrates the need to pinpoint the geographical location of a complete image based on only a portion of it, as demonstrated in 6(a). As illustrated in Figure 6(b), all agents recognized the Statue of Liberty in Figure 6(a), and some identified the presence of part of the Eiffel Tower at the edge of the picture. For instance, GPT-4V concluded that the buildings in these two locations appeared in the same image. However, as is known through the knowledge of other agents (Gemini), a scaled-down version of the Statue of Liberty has been erected on Swan Island, an artificial island in the Seine River in France. By marking both the Eiffel Tower and the island on the Open Street Map (OSM) manually, as shown in Figure 6(c), it is evident that they are merely 1.3 kilometers apart in a straight line. By utilizing the proposed framework, agents discuss and summarize the location depicted in the picture to be Paris, France, as shown in Figure 6(d). Thus, without human intervention, this framework demonstrates the effectiveness of doing geo-localization tasks.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our work proposes a swarm intelligence geo-localization framework, smileGeo, which contains the process of the review mechanism in agent discussions along with a dynamic learning strategy of agent collaboration social network, to achieve open-world geo-localization tasks. In addition, we construct a novel geo-localization dataset, GeoGlobe for evaluation and it will be public. All of the contributions we claimed in both abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: At present, the LVLM agent framework we proposed can only search the Internet autonomously. Our agent still has shortcomings in the use of other multiple tools. We stated in our future outlook that our follow-up work will solve this problem.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This work is a solution to the problem of geo-localization in application scenarios. We have provided the source code and will release the related dataset, as the dataset is relatively large (about 32 GB) and cannot be uploaded as an attachment.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided the source code and will release the related dataset once the paper is accepted, as the dataset is relatively large (about 32 GB) and cannot be uploaded as an attachment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the anonymous code link: https://anonymous.4open.science/r/ViusalGeoLocalization-F8F5/. In this link, we also provide a small-scale dataset we collected for people to reproduce the results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We explain all the settings in both the main paper (Experiments) and the appendix (Implementation Details).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We deploy a two-sided t-test with $p < 0.05$ for our baseline experiments.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We announce the compute resources in the appendix (Implementation Details).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The codes used in our paper are all open source, and the data used in the paper come from copyright-free images on the Internet.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We have an outlook on our research in the section Conclusion, which can be widely used in robot positioning and navigation in the future.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: The data sets we collect have been manually reviewed twice, and all data containing various types of sensitive information or copyright risks have been filtered out.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We list and acknowledge all other open-source codes we used in the file 'README.md' and we follow the license for existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: In this paper, we provide the algorithm of the code and introduce the dataset in detail (in the appendix).

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper aims to address visual geo-localization tasks and does not contain any experiments with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not contain any experiments with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.