

GraphGeo: Multi-Agent Debate Framework for Visual Geo-localization with Heterogeneous Graph Neural Networks

Heng Zheng¹, Yuling Shi², Xiaodong Gu², Haochen You³, Zijian Zhang⁴, Lubin Gan⁵, Hao Zhang⁶, Wenjun Huang⁷, and Jin Huang^{1,*}

¹South China Normal University, 2024025450@m.scnu.edu.cn,
huangjin@m.scnu.edu.cn

²Shanghai Jiao Tong University, yuling.shi@sjtu.edu.cn,
xiaodong.gu@sjtu.edu.cn

³Columbia University, hy2854@columbia.edu

⁴University of Pennsylvania, zzjharry@umich.edu

⁵University of Science and Technology of China, ganlubin@mail.ustc.edu.cn

⁶University of Chinese Academy of Sciences, zhang_hao1999@yeah.net

⁷Sun Yat-sen University, huangwj98@mail2.sysu.edu.cn

*Corresponding author

Abstract. Visual geo-localization requires extensive geographic knowledge and sophisticated reasoning to determine image locations without GPS metadata. Traditional retrieval methods are constrained by database coverage and quality. Recent Large Vision-Language Models (LVLMs) enable direct location reasoning from image content, yet individual models struggle with diverse geographic regions and complex scenes. Existing multi-agent systems improve performance through model collaboration but treat all agent interactions uniformly. They lack mechanisms to handle conflicting predictions effectively. We propose **GraphGeo**, a multi-agent debate framework using heterogeneous graph neural networks for visual geo-localization. Our approach models diverse debate relationships through typed edges, distinguishing supportive collaboration, competitive argumentation, and knowledge transfer. We introduce a dual-level debate mechanism combining node-level refinement and edge-level argumentation modeling. A cross-level topology refinement strategy enables co-evolution between graph structure and agent representations. Experiments on multiple benchmarks demonstrate GraphGeo significantly outperforms state-of-the-art methods. Our framework transforms cognitive conflicts between agents into enhanced geo-localization accuracy through structured debate.

Keywords: Large Visual Geo-localization · Multi-agent · Debate · Heterogeneous Graph Neural Networks · Large Vision-Language Models.

1 Introduction

Visual geo-localization involves identifying the geographic location of an image without relying on metadata like GPS tags. This task serves an important role

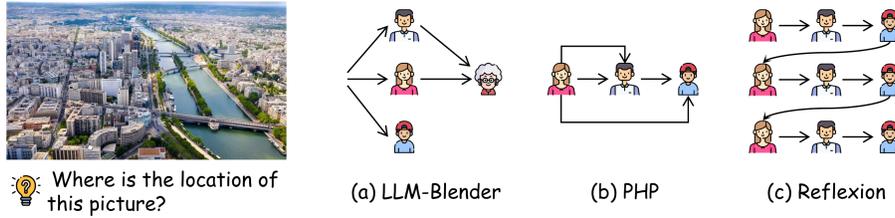


Fig. 1: Comparison of multi-agent collaboration frameworks for visual geolocalization. (a) LLM-Blender: agents independently predict and aggregate through simple voting, lacking conflict resolution. (b) PHP: sequential collaboration with fixed order, prone to error propagation. (c) Reflexion: parallel self-reflection without inter-agent debate. These methods treat agent interactions uniformly and cannot effectively resolve conflicting predictions through structured argumentation.

in various applications, including analyzing historical human mobility, assisting autonomous navigation, and providing location-aware recommendation services [14, 4]. Traditional methods tackle this problem by retrieving similar images from large geo-tagged databases [3, 26]. However, their performance is limited by database scope, update frequency, and record accuracy. Recent progress in Large Vision-Language Models offers new opportunities by enabling direct location reasoning through image content [32, 11]. These models excel in interpreting complex visual information and incorporate extensive geospatial knowledge from pre-training on diverse multi-modal data.

Despite their impressive capabilities, individual LVLMs encounter challenges in complex geo-localization scenarios [11]. A single model may struggle with fine-grained features across culturally diverse regions and show preference for well-represented training locations. These models produce inconsistent predictions when facing ambiguous visual cues or specialized geographic scenes. Researchers have developed multi-agent collaboration frameworks combining multiple LVLMs with complementary strengths [38, 15, 20]. Current systems typically employ independent exploration with simple result aggregation or sequential collaboration with predetermined interaction orders [11]. While these strategies leverage model diversity for improved performance, they lack robust mechanisms for resolving conflicting predictions. Simple aggregation cannot determine prediction reliability or synthesize competing interpretations effectively. These approaches miss opportunities for structured argumentation to expose reasoning flaws and refine predictions through iterative feedback [20].

The core challenge lies in recognizing the heterogeneous nature of agent interactions in multi-agent systems. Different agents analyzing the same image form distinct relationships based on their prediction consistency and confidence levels. Agents with similar predictions benefit from supportive collaboration to consolidate shared understanding and strengthen correct reasoning. In contrast, agents

producing conflicting predictions require competitive argumentation to challenge assumptions and identify errors through critical examination. Real-world geo-localization often presents such conflicts. Architectural features might indicate one region, yet natural landscape elements suggest another location entirely. Simple aggregation methods like majority voting or confidence-weighted averaging fail to determine the reliability of different predictions or recognize when apparent conflicts actually reveal complementary evidence. Knowledge should flow directionally from high-confidence agents to uncertain ones for guided refinement. Human experts resolve disagreements by articulating their reasoning, presenting supporting evidence, challenging alternative hypotheses, and progressively building consensus through structured debate. This process exposes reasoning flaws and synthesizes diverse perspectives into comprehensive understanding. Traditional graph neural networks assume homogeneous connectivity and apply identical message passing functions across all edges. This approach misrepresents the semantic diversity of multi-agent interactions. Supportive consolidation, competitive critique, and unidirectional knowledge transfer require different treatment to enable appropriate information exchange. Current systems cannot leverage the complementary benefits of cooperation and competition without explicitly distinguishing these interaction types.

Beyond recognizing the heterogeneous nature of interactions, another critical challenge emerges from the dynamic evolution of debate relationships. Multi-agent geo-localization demands modeling collaboration as diverse debate rather than uniform information exchange. Agent interactions serve three distinct purposes requiring different computational mechanisms. When predictions align, agents should engage in supportive collaboration to consolidate shared understanding by reinforcing consistent evidence. When predictions conflict, agents need competitive argumentation to challenge hypotheses and expose reasoning gaps. When confidence levels differ substantially, knowledgeable agents should transfer expertise to uncertain agents through guided reasoning. Existing graph-based multi-agent systems apply identical message passing functions across all connections. This uniform treatment fails to capture the semantic richness of these fundamentally different interaction modes. Furthermore, debate relationships shift dynamically as discussions progress. Two agents initially disagreeing may converge toward consensus through sustained argumentation. Conversely, initially aligned agents may discover subtle differences requiring deeper examination. Standard graph neural networks update node representations yet leave edge relationships static or only implicitly modified through node state changes. This design cannot explicitly track how argumentation dynamics between specific agent pairs evolve across debate rounds. The interaction history between agents carries valuable signals about argument strength, persuasiveness, and reliability. Capturing this temporal evolution of pairwise debate states requires explicit modeling beyond node-level representation updates.

To address these challenges, we propose GraphGeo, a multi-agent debate framework for visual geo-localization based on heterogeneous graph neural networks. Our framework transforms cognitive conflicts into valuable resources

through structured debate mechanisms. GraphGeo models agent collaboration as structured debate over a heterogeneous graph with nodes representing LVLM agents and edges encoding different debate relationships. We design three edge types: agreement edges for knowledge consolidation through mutual reinforcement, conflict edges for competitive argumentation through critical feedback, and transfer edges for guided knowledge sharing from confident to uncertain agents. We introduce a dual-level debate mechanism. At the node level, agents refine representations by aggregating relation-specific messages through tailored message passing. At the edge level, learnable edge representations explicitly model pairwise argumentation states and capture debate dynamics. These edge states evolve via recurrent updates and influence subsequent node reasoning. Furthermore, we develop cross-level topology refinement enabling bidirectional influence between graph structure and node representations. In summary, our contributions are as follows:

- We identify that existing multi-agent geo-localization systems treat agent interactions homogeneously, overlooking the diverse nature of debate relationships and limiting effective conflict resolution among agents.
- We propose GraphGeo, a novel framework using heterogeneous graph neural networks with a dual-level debate mechanism and cross-level topology refinement, enabling structured argumentation and adaptive collaboration patterns.
- Extensive experiments demonstrate GraphGeo achieves substantial improvements over state-of-the-art methods, with ablation studies validating each component’s effectiveness.

2 Related Works

2.1 Visual Geo-localization Methods

Visual geo-localization aims to determine the geographic location of an image using only visual content without relying on GPS metadata [14]. Traditional approaches tackle this problem through large-scale image retrieval from geo-tagged databases [39, 29]. NetVLAD [3] pioneered end-to-end trainable architectures by introducing a generalized VLAD layer inspired by Vector of Locally Aggregated Descriptors. Recent methods like CosPlace [6] reformulate the task as classification to enable training on larger datasets, while two-stage strategies such as Patch-NetVLAD [13] combine global descriptors for initial ranking with local patch-level matching for refinement. However, retrieval-based systems remain constrained by database coverage and quality, with their applicability often limited by available knowledge and regional data protection laws. Recent progress in Vision-Language Models has introduced a new paradigm enabling direct location reasoning from images [28, 18]. Models like CLIP [28] and GeoCLIP [32] incorporate semantic reasoning into feature representations, while generative VLMs such as GPT-4V [22] and Gemini [31] demonstrate the ability to infer geographic locations directly. GeoReasoner [12] addresses data scarcity by creating highly

locatable street-view datasets and integrating human inference knowledge from real geo-localization games. However, existing evaluations remain limited with most work examining only a handful of models, leaving critical factors such as model family, scale, and failure cases underexplored. Single LVLM approaches encounter challenges with fine-grained features across culturally diverse regions and produce inconsistent predictions when facing ambiguous visual cues or specialized geographic scenes [11].

2.2 Multi-agent Framework for LLM/LVLMs

Large language models have evolved from single-agent systems to collaborative multi-agent frameworks for complex task solving [24, 17]. Multi-agent systems deploy specialized agents with complementary capabilities to overcome limitations of individual models [38]. AutoGen [38] provides a high-level interface for orchestrating conversations between multiple agents through specified prompts and tool access. MetaGPT [15] integrates human workflows into multi-agent collaborations by incorporating standardized operating procedures to reduce errors. Multi-agent debate frameworks leverage structured argumentation to improve reasoning quality beyond single-agent approaches [9, 20]. MAD [20] introduces tit-for-tat debate states to encourage divergent thinking and correct distorted reasoning among agents. Agent4Debate [40] employs specialized agents for searching, analyzing, writing, and reviewing throughout competitive debate processes. Graph neural networks provide natural representations for multi-agent interactions by modeling agents as nodes and relationships as edges [16, 1]. MAGEC [10] combines GNNs with multi-agent reinforcement learning to capture complex cooperative and competitive dynamics. Heterogeneous graph attention mechanisms enable agents to adaptively weight information from different neighbor types [37, 30]. Graph-based MARL achieves scalability for large-scale systems through message-passing and neighbor sampling strategies [19, 35]. Swarm intelligence frameworks smileGeo [11] aggregate multiple LVLM agents for improved geo-localization through collaborative reasoning. Current systems typically employ independent exploration with simple result aggregation or sequential collaboration with predetermined orders. These approaches lack explicit modeling of heterogeneous interaction types and fail to distinguish between supportive collaboration, competitive argumentation, and knowledge transfer relationships among agents.

3 Preliminary

3.1 Problem Formulation

Visual geo-localization aims to predict the geographic location \mathcal{Y} of a given image \mathcal{X} without relying on GPS metadata. We formulate this task within a multi-agent framework where N large vision-language models serve as collaborative agents $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$. Each agent v_i generates location predictions \mathcal{Y}_{v_i} with confidence scores $c_{v_i} \in [0, 1]$. The collaborative prediction process can be modeled

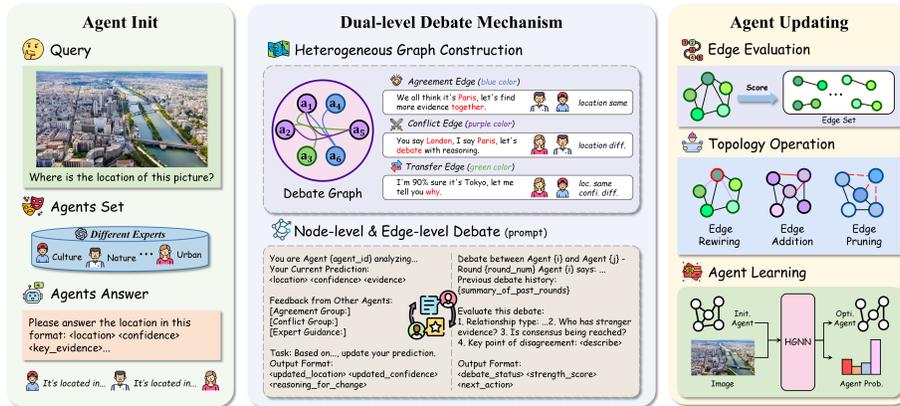


Fig. 2: The framework overview of GraphGeo. It contains a dual-level debate mechanism and cross-level topology refinement for visual geo-localization. The first part constructs a heterogeneous graph with three edge types to model diverse agent interactions. The second part employs node-level and edge-level debate to refine predictions through structured argumentation. The third part optimizes graph topology by evaluating edge utility and adjusting connections during multi-round debates.

as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where nodes represent agents and edges \mathcal{E} encode their interaction patterns. The optimal prediction \mathcal{Y}^* emerges from aggregating diverse agent perspectives weighted by their expertise and mutual influence within the graph structure.

3.2 Graph Neural Networks for Agent Collaboration

Graph neural networks provide a principled framework for learning representations through iterative message passing. Given the agent collaboration graph \mathcal{G} , each agent v_i maintains a hidden state $\mathbf{h}_i^{(l)} \in \mathbb{R}^d$ at layer l . The standard message passing mechanism updates node representations by aggregating neighborhood information through

$$\mathbf{h}_i^{(l+1)} = \text{UPDATE}^{(l)} \left(\mathbf{h}_i^{(l)}, \text{AGGREGATE}^{(l)} \left(\{\mathbf{m}_{ji}^{(l)} \mid v_j \in \mathcal{N}(v_i)\} \right) \right), \quad (1)$$

where $\mathcal{N}(v_i)$ denotes the neighbors of agent v_i and $\mathbf{m}_{ji}^{(l)}$ represents the message from agent v_j to v_i at layer l . However, this homogeneous treatment of agent interactions fails to capture the competitive and collaborative dynamics essential for effective geo-localization. Agents may produce conflicting predictions that require resolution through argumentation rather than simple averaging, and different images demand different collaboration patterns among agents.

4 Methodology

4.1 Heterogeneous Graph Construction with Debate Edges

We construct a heterogeneous agent collaboration graph that explicitly models diverse interaction types through typed edges. The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$ incorporates a set of relation types $\mathcal{R} = \{r_{\text{agree}}, r_{\text{conflict}}, r_{\text{transfer}}\}$ to distinguish between different debate mechanisms. For each image \mathcal{X} , agent nodes $v_i \in \mathcal{V}$ have features $\mathbf{x}_i^{(0)}$ that concatenate the VAE-encoded image representation with learnable agent embeddings $\mathbf{e}_i \in \mathbb{R}^{d_e}$. The edge set \mathcal{E} is dynamically constructed based on initial agent predictions. We compute the geodesic distance $d_{\text{geo}}(\mathcal{Y}_{v_i}, \mathcal{Y}_{v_j})$ between predicted locations and assign relation types through

$$r_{ij} = \begin{cases} r_{\text{agree}} & \text{if } d_{\text{geo}}(\mathcal{Y}_{v_i}, \mathcal{Y}_{v_j}) < \tau_{\text{agree}} \\ r_{\text{conflict}} & \text{if } d_{\text{geo}}(\mathcal{Y}_{v_i}, \mathcal{Y}_{v_j}) > \tau_{\text{conflict}} \\ r_{\text{transfer}} & \text{if } c_{v_i} - c_{v_j} > \tau_{\text{conf}} \end{cases}, \quad (2)$$

where τ_{agree} , τ_{conflict} , and τ_{conf} are threshold hyperparameters. Agreement edges connect agents with consistent predictions to facilitate knowledge consolidation, conflict edges link agents with divergent views to enable competitive debate, and transfer edges direct information flow from high-confidence to low-confidence agents.

4.2 Competitive Message Passing with Node-level Debate

At each layer l , we replace standard message aggregation with a competitive debate mechanism. For each agent v_i , we partition its neighbors $\mathcal{N}(v_i)$ according to edge types as $\mathcal{N}_r(v_i) = \{v_j \mid (v_j, v_i, r) \in \mathcal{E}\}$ for $r \in \mathcal{R}$. The node-level debate generates relation-specific messages tailored to the interaction type. For agreement edges, agent v_j generates supportive messages that reinforce shared understanding through

$$\mathbf{m}_{ji}^{\text{agree}} = \mathbf{W}_{\text{agree}} \cdot \sigma(\mathbf{h}_j^{(l)} \odot \mathbf{h}_i^{(l)}), \quad (3)$$

where $\mathbf{W}_{\text{agree}} \in \mathbb{R}^{d \times d}$ is a learnable weight matrix and \odot represents element-wise multiplication. For conflict edges, agent v_j generates critical messages that challenge the current hypothesis through

$$\mathbf{m}_{ji}^{\text{conflict}} = \mathbf{W}_{\text{conflict}} \cdot \text{ReLU}(\mathbf{h}_j^{(l)} - \mathbf{h}_i^{(l)}) \quad (4)$$

Transfer edges employ an attention-based mechanism where high-confidence agent v_j selectively transfers relevant knowledge through

$$\mathbf{m}_{ji}^{\text{transfer}} = \alpha_{ji} \mathbf{W}_{\text{transfer}} \mathbf{h}_j^{(l)}, \quad \alpha_{ji} = \frac{\exp(c_{v_j} \cdot \mathbf{h}_j^{(l)\top} \mathbf{W}_{\text{attn}} \mathbf{h}_i^{(l)})}{\sum_{v_k \in \mathcal{N}_{\text{transfer}}(v_i)} \exp(c_{v_k} \cdot \mathbf{h}_k^{(l)\top} \mathbf{W}_{\text{attn}} \mathbf{h}_i^{(l)})}, \quad (5)$$

where α_{ji} weights the transfer by both confidence c_{v_j} and semantic relevance. The agent v_i then aggregates these heterogeneous messages through relation-aware pooling

$$\mathbf{m}_i^{(l)} = \bigoplus_{r \in \mathcal{R}} \text{POOL}(\{\mathbf{m}_{ji}^r \mid v_j \in \mathcal{N}_r(v_i)\}), \quad (6)$$

where \bigoplus denotes a learnable combination operator that balances different debate mechanisms.

4.3 Edge-level Debate through Cross-agent Interaction

Beyond node-level refinement, we introduce edge-level debate that explicitly models pairwise interactions to capture fine-grained argumentation dynamics. For each edge $(v_i, v_j, r) \in \mathcal{E}$, we construct a debate representation $\mathbf{d}_{ij}^{(l)}$ that encodes the interaction state between agents. This edge representation evolves through

$$\mathbf{d}_{ij}^{(l)} = \mathbf{W}_{\text{edge}}^r \cdot [\mathbf{h}_i^{(l)} \oplus \mathbf{h}_j^{(l)} \oplus \Delta\mathcal{Y}_{ij}], \quad \mathbf{d}_{ij}^{(l+1)} = \text{GRU}(\mathbf{d}_{ij}^{(l)}, \mathbf{h}_i^{(l+1)} + \mathbf{h}_j^{(l+1)}), \quad (7)$$

where $\mathbf{W}_{\text{edge}}^r \in \mathbb{R}^{d_e \times 3d}$ is a relation-specific projection matrix and GRU maintains debate history while incorporating updated node states. The edge representations subsequently influence node updates through edge-to-node message passing

$$\mathbf{m}_i^{\text{edge},(l)} = \sum_{v_j \in \mathcal{N}(v_i)} \beta_{ij} \mathbf{d}_{ij}^{(l)}, \quad \beta_{ij} = \text{softmax}_j(\mathbf{d}_{ij}^{(l)\top} \mathbf{W}_{\text{attend}} \mathbf{h}_i^{(l)}) \quad (8)$$

The final node update integrates both node-level and edge-level messages through a gated fusion mechanism

$$\mathbf{h}_i^{(l+1)} = \text{GRU}(\mathbf{h}_i^{(l)}, \gamma_i \mathbf{m}_i^{(l)} + (1 - \gamma_i) \mathbf{m}_i^{\text{edge},(l)}), \quad (9)$$

where the gate $\gamma_i = \sigma(\mathbf{w}_\gamma^\top [\mathbf{h}_i^{(l)} \oplus \mathbf{m}_i^{(l)} \oplus \mathbf{m}_i^{\text{edge},(l)}])$ adaptively balances the two information sources.

4.4 Cross-level Topology Refinement

To enable dynamic adaptation of the debate structure, we introduce a cross-level topology refinement mechanism. After L layers of message passing, we obtain final agent representations $\mathbf{H} = \{\mathbf{h}_i^{(L)} \mid v_i \in \mathcal{V}\}$ that encode refined predictions. These representations feedback to the graph topology through a learned edge scoring function

$$s_{ij} = \text{MLP}_{\text{score}}([\mathbf{h}_i^{(L)} \oplus \mathbf{h}_j^{(L)} \oplus \mathbf{d}_{ij}^{(L)} \oplus \Delta\mathbf{h}_{ij}]), \quad (10)$$

where $\Delta\mathbf{h}_{ij} = \mathbf{h}_i^{(L)} - \mathbf{h}_i^{(0)}$ captures the representational change induced by debate. Edges with low utility scores are pruned while new edges are added between

agents whose representations become similar after debate. The topology learning objective combines task-specific loss with graph regularization

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{geo}} + \lambda_{\text{sparse}} \sum_{(i,j) \in \mathcal{E}} s_{ij} + \lambda_{\text{diverse}} \mathcal{H}(\mathcal{R}), \quad (11)$$

where \mathcal{L}_{geo} measures the geodesic distance between predicted and ground-truth locations, the sparsity term penalizes excessive edge connectivity, and $\mathcal{H}(\mathcal{R})$ represents the entropy over relation type distribution. This cross-level interaction ensures that the topology evolves to support effective collaboration patterns while node representations adapt to leverage the refined debate structure.

5 Experiments

5.1 Datasets

We evaluate our GraphGeo framework on four widely-adopted geo-localization benchmarks. The **IM2GPS3K** dataset [33] contains 3,000 images randomly sampled from the original IM2GPS collection, providing diverse geographic coverage with images spanning multiple continents and representing various scenes including urban landscapes, natural environments, and landmarks. Another is the **YFCC4K** dataset [33], derived from the Yahoo Flickr Creative Commons 100 Million collection, consists of 4,536 geo-tagged images with broad visual diversity and geographic distribution. For comprehensive evaluation, we utilize the newly constructed **GeoGlobe** dataset [11], which divides images into two categories: **GeoGlobe (Natural)** containing natural attractions and **GeoGlobe (ManMade)** featuring man-made landmarks. This dataset addresses limitations of existing benchmarks by filtering out ambiguous images and providing clearer localization targets, while offering substantial coverage across different geographic regions and environmental conditions.

5.2 Baselines

We compare GraphGeo against three categories of baselines. First, we evaluate **single LVLM agents** including both open-source models (Qwen-VL [5], Qwen2-VL [34], InternVL2 [8], LLaVA series [21], CogVLM [36]) and proprietary models (Claude-3.5-Sonnet [2], GPT-4 series [23], Gemini series [31]) to establish individual model performance. Second, we include **traditional visual geo-localization methods** such as NetVLAD [3], which pioneered trainable VLAD pooling, GeM [27] using generalized mean pooling, CosPlace [7] reformulating the task as classification for scalability, Translocator [25], PIGEON, StreetCLIP, and GeoCLIP [32] integrating semantic reasoning into features. Third, we compare with **smileGeo** [11], a swarm intelligence framework aggregating multiple LVLM agents through collaborative reasoning, representing the current state-of-the-art in multi-agent geo-localization.

Table 1: Results of different single LVLB baselines and comparison with baselines. Bold indicates the statistically significant improvements over the best baseline.

Model	IM2GPS3K	YFCC4K	GeoGlobe (Natural)	GeoGlobe (ManMade)
<i>Open-source LVLBs</i>				
Infi-MM	14.7000	6.5256	19.2020	21.4145
Qwen-VL	32.4667	14.3959	42.3940	37.4556
Qwen2-VL-7B	36.2333	15.8724	45.6783	52.3142
InternVL2-8B	33.5667	14.9213	40.1256	48.7623
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
CogVLM2	28.4333	12.5678	35.4521	42.3187
<i>Proprietary LVLBs</i>				
Claude-3-Opus	23.8333	12.3457	33.1671	40.6954
Claude-3.5-Sonnet	41.2667	18.6735	55.3214	76.8456
GPT-4V	43.5333	19.2147	58.4672	80.1234
GPT-4o	46.8333	20.4569	64.7823	85.2617
GPT-4o-mini	45.0000	18.3422	62.0948	84.5911
Gemini-1.5-Flash	39.4667	17.2341	54.8912	74.5623
Gemini-1.5-Pro	47.3667	19.9956	62.3441	82.8131
<i>Visual Geo-localization Methods</i>				
NetVLAD	16.6303	7.4876	26.5134	28.9955
GeM	14.4907	6.5243	23.1022	25.4175
CosPlace	17.6686	7.9551	28.1688	30.2782
Translocator	31.0978	13.4039	26.1776	34.1971
PIGEON	29.5632	12.8914	35.7891	41.2365
StreetCLIP	33.2145	14.6782	39.5621	44.8912
GeoCLIP	34.4728	15.1719	38.2519	45.9174
smileGeo	47.7667	21.5168	76.0599	85.4603
GraphGeo (Ours)	49.1456	22.1623	77.8217	87.9341

5.3 Metrics

Following standard practice in visual geo-localization, we adopt **street-level accuracy** as our primary evaluation metric. This metric measures the percentage of images correctly localized within a given distance threshold from the ground truth location. Specifically, we report accuracy across multiple granularity levels ranging from fine-grained street-level (1 km) to coarse-grained continent-level (2500 km) thresholds. Street-level accuracy directly reflects the practical utility of geo-localization systems in real-world applications such as navigation and

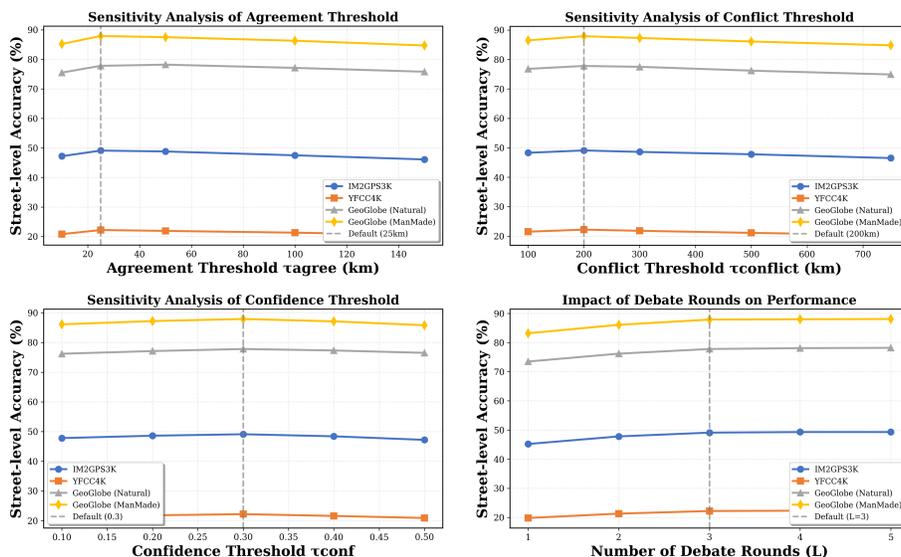


Fig. 3: Hyperparameter sensitivity analysis. (a) Agreement threshold τ_{agree} shows optimal values vary across datasets due to different geographic distributions, with default at 25km. (b) Conflict threshold $\tau_{conflict}$ balances critical debate and noise, where lower values create more conflict edges. (c) Confidence threshold τ_{conf} controls knowledge transfer between agents, balancing information flow and over-reliance. (d) Debate rounds show performance gains diminish after 3 rounds, making $L = 3$ optimal for accuracy-cost trade-off.

location-aware services. For each dataset, we compute the proportion of test images whose predicted locations fall within the specified radius of their true geographic coordinates. This metric provides an intuitive and interpretable assessment of localization performance while enabling fair comparison across different methods and datasets.

5.4 Implementation Details

We implement GraphGeo all experiments on NVIDIA A100 GPUs. Our multi-agent system incorporates a diverse set of LVLm agents selected from both open-source and proprietary models to ensure complementary reasoning capabilities. For graph construction, we initialize the heterogeneous graph with typed edges based on prediction similarity, setting distance thresholds $\tau_{agree} = 25\text{km}$, $\tau_{conflict} = 200\text{km}$, and $\tau_{conf} = 0.3$ for edge type assignment. The dual-level debate mechanism operates through $L = 3$ layers of message passing with hidden dimension $d = 512$. We employ the Adam optimizer with learning rate 1×10^{-4} and train for 50 epochs with batch size 32. The cross-level topology refinement uses hyperparameters $\lambda_{\text{sparse}} = 0.01$ and $\lambda_{\text{diverse}} = 0.05$ to balance prediction

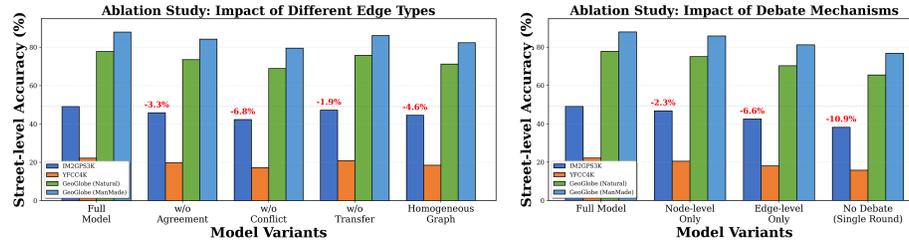


Fig. 4: Ablation Study on Edge Types and Debate Mechanism Caption: Ablation studies on heterogeneous edge types and dual-level debate mechanism. Conflict edges are most critical, with 6.8% performance drop when removed, highlighting the essential role of competitive debate in error correction. Node-level aggregation proves more critical than edge-level modeling, while removing the entire debate mechanism causes 10.9% performance drop on IM2GPS3K, validating the core value of iterative multi-agent argumentation.

accuracy with graph structure quality. During inference, we aggregate final agent predictions through confidence-weighted voting, where each agent’s contribution is determined by its refined confidence score after multi-round debate.

5.5 Main Result

Table 1 presents our comprehensive evaluation against various baselines including open-source LVLMs, proprietary models, traditional geo-localization methods, and the state-of-the-art multi-agent system smileGeo. GraphGeo achieves the best performance across all four benchmarks, with street-level accuracy reaching 49.1% on IM2GPS3K and 87.9% on GeoGlobe ManMade. Our framework outperforms the strongest baseline smileGeo by notable margins on both IM2GPS3K and GeoGlobe datasets. The core insight of GraphGeo is that agent interactions are inherently heterogeneous and require differentiated treatment through typed edges. By modeling collaboration as agreement edges, competition as conflict edges, and guidance as transfer edges, our framework enables agents to engage in structured debate rather than simple aggregation. The consistent improvements across diverse geographic distributions validate that heterogeneous graph modeling captures the semantic richness of multi-agent argumentation, transforming cognitive conflicts into enhanced geo-localization accuracy through principled debate mechanisms.

5.6 Model Analysis

Our analysis reveals distinct performance patterns across model categories. Proprietary LVLMs demonstrate strong individual capabilities, while open-source models show more variability in performance. Traditional geo-localization methods lag significantly behind vision-language approaches, confirming the value of

semantic reasoning for this task. The multi-agent baseline smileGeo surpasses all individual models through collaborative reasoning, yet our GraphGeo framework achieves further improvements. This progression demonstrates a fundamental principle that structured argumentation through heterogeneous graph modeling drives optimal performance. Simple agent ensemble treats all interactions uniformly and cannot resolve conflicts effectively, while our approach explicitly distinguishes between supportive collaboration, competitive argumentation, and knowledge transfer. The dual-level debate mechanism enables both coarse-grained message aggregation at the node level and fine-grained interaction modeling at the edge level, allowing agents to internalize diverse perspectives while maintaining explicit debate dynamics that evolve across multiple rounds of refinement.

5.7 Hyper-parameter Analysis

Figure 3 examines the sensitivity of key hyperparameters across our four benchmarks. The agreement threshold shows optimal values vary by dataset, with 25km working best overall as it balances supportive collaboration without over-connecting dissimilar predictions. The conflict threshold at 200km effectively separates genuine disagreements from noise. Lower values create excessive conflict edges that dilute critical debate signals. The confidence threshold provides the right balance for knowledge transfer at 0.3, preventing over-reliance on potentially incorrect predictions while still enabling guidance flow. Debate rounds show performance gains through three iterations, after which improvements plateau. The dynamic nature of our graph topology allows the framework to adapt edge connections based on debate effectiveness, with cross-level refinement enabling bidirectional influence between structure and representations. This adaptive mechanism ensures that the most productive debate relationships strengthen over time while unproductive connections are pruned, creating an efficient collaboration structure tailored to each specific image and its geographic ambiguity.

5.8 Ablation Study

Figure 4 systematically evaluates each component’s contribution to our framework. Removing conflict edges causes the largest performance drop of 6.8% on IM2GPS3K, highlighting that competitive argumentation is essential for error correction through critical debate. Agreement edges contribute 3.3% and transfer edges add 1.9%, while the homogeneous graph variant loses 4.6% by treating all interactions uniformly. For debate mechanisms, eliminating all iterative refinement causes a dramatic 10.9% performance drop, validating that multi-round argumentation forms the core value of our approach. The results reveal a key insight that different agent relationships serve fundamentally different purposes in collaborative geo-localization. When agents agree, they need mechanisms to consolidate shared understanding and reinforce correct reasoning. When agents

conflict, they require structured channels for challenging assumptions and exposing reasoning flaws. When confidence levels differ, knowledge should flow directionally from certain to uncertain agents. Our heterogeneous graph design captures these distinct interaction semantics, enabling agents to leverage both cooperation and competition for enhanced reasoning quality.

6 Conclusion

This paper introduces GraphGeo, a multi-agent debate framework that transforms visual geo-localization through heterogeneous graph neural networks. By explicitly modeling diverse agent interactions as typed edges and implementing dual-level debate mechanisms, our framework enables structured argumentation that effectively resolves conflicting predictions. The cross-level topology refinement strategy allows dynamic adaptation of collaboration patterns, turning cognitive conflicts into enhanced localization accuracy. GraphGeo advances beyond simple aggregation approaches, demonstrating that heterogeneous interaction modeling and structured debate are crucial for multi-agent systems. This work opens new directions for applying graph-based reasoning to collaborative AI tasks requiring consensus under uncertainty.

References

1. Agarwal, A., Kamath, U., Jindal, R.: Learning to communicate with deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems* **32** (2019)
2. Anthropic: The claude 3 model family: Opus, sonnet, haiku. <https://www.anthropic.com/claude> (2024)
3. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5297–5307 (2016)
4. Baatz, G., Saurer, O., Köser, K., Pollefeys, M.: Large scale visual geo-localization of images in mountainous terrain. In: *Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*. pp. 517–530. Springer (2012)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. In: *arXiv preprint arXiv:2308.12966* (2023)
6. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4878–4888 (2022)
7. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4878–4888 (2022)
8. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with model scaling. *arXiv preprint arXiv:2410.12525* (2024)

9. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325 (2023)
10. Goeckner, A., Sui, Y., Martinet, N., Li, X., Zhu, Q.: Graph neural network-based multi-agent reinforcement learning for resilient distributed coordination of multi-robot systems. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8586–8593. IEEE (2024)
11. Han, X., Zhu, C., Zhao, X., Zhu, H.: Swarm intelligence in geo-localization: A multi-agent large vision-language model collaborative framework. arXiv preprint arXiv:2408.11312 (2024)
12. Han, X., Zhu, C., Zhao, X., Zhu, H.: GeoReasoner: Enhancing geospatial reasoning in large vision language models. arXiv preprint arXiv:2407.12345 (2024)
13. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14141–14152 (2021)
14. Hays, J., Efros, A.A.: im2gps: estimating geographic information from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2008)
15. Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., Schmidhuber, J.: MetaGPT: Meta programming for a multi-agent collaborative framework. In: The Twelfth International Conference on Learning Representations (2024)
16. Jiang, J., Dun, C., Huang, T., Lu, Z.: Graph convolutional reinforcement learning. In: International Conference on Learning Representations (2020)
17. Li, G., Hammoud, H.A.A.K., Itani, H., Khizbullin, D., Ghanem, B.: CAMEL: Communicative agents for “mind” exploration of large language model society. *Advances in Neural Information Processing Systems* **36** (2023)
18. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning* pp. 12888–12900 (2022)
19. Li, Q., Gama, F., Ribeiro, A., Prorok, A.: Graph neural networks for decentralized multi-robot path planning. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 11785–11792. IEEE (2020)
20. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., Tu, Z.: Encouraging divergent thinking in large language models through multi-agent debate. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 17889–17904 (2024)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in Neural Information Processing Systems* **36**, 34892–34916 (2023)
22. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
23. OpenAI: Gpt-4 technical report (2023)
24. Park, J.S., O’Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442 (2023)
25. Pramanick, S., Bhowmik, A., et al.: Translocator: Transformer-based multi-task learning framework for large-scale visual geo-localization. arXiv preprint arXiv:2205.11097 (2022)
26. Radenović, F., Tolias, G., Chum, O.: Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(7), 1655–1668 (2018)

27. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 41, pp. 1655–1668 (2018)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. pp. 8748–8763. PMLR (2021)
29. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1–7. IEEE (2007)
30. Seraj, E., Silva, A., Gombolay, M.: Heterogeneous graph attention networks for scalable multi-robot scheduling with temporospatial constraints. In: *Autonomous Robots*. vol. 45, pp. 937–957 (2021)
31. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Ahn, L.B.E., Aguilar, J., Cheng, Kree Cole-McLaughlin Biering, J.D., Hwang, o.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024)
32. Vivek, C., Khurana, G.S., Jain, Y., Vaidya, K., Mehta, V., et al.: Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems* **36** (2024)
33. Vo, N., Jacobs, N., Hays, J.: Revisiting im2gps in the deep learning era. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 2621–2630 (2017)
34. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024)
35. Wang, T., Liao, R., Ba, J., Fidler, S.: NerveNet: Learning structured policy with graph neural networks. In: *International Conference on Learning Representations* (2018)
36. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079* (2023)
37. Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., Yu, P.S.: Heterogeneous graph attention network. In: *The World Wide Web Conference*. pp. 2022–2032 (2019)
38. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., Wang, C.: Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. In: *arXiv preprint arXiv:2308.08155* (2023)
39. Zamir, A.R., Shah, M.: Accurate geo-localization of street-views using location-oriented information from the web. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2010)
40. Zhao, X., Wang, X., Chen, W., Liu, Y., Zhang, H.: Agent4debate: A multi-agent framework for automatic debate generation. *arXiv preprint arXiv:2401.09357* (2024)