EMORY

Undergraduate Research Programs

Pathways Center

GeoLens: A Temporal-aware, Multimodal Retrieval-Augmented System for Geospatial Question Answering

Mingyang Li, Henry Ning, Ethan Yang

Department of Computer Science, Emory University

Abstract/Intro/Motivation

Geospatial Question Answering is central to everyday. However, existing Geographic Information Systems lack natural language understanding while Large Language Models (LLMs) struggles with spatial grounding. Although recent Retrieval-Augmented Generation (RAG) approaches have made progress by connecting LLMs with external databases, most approaches remain text-only and struggle to handle multimodal signals. Therefore, we presents **GeoLens**, a multimodal RAG framework that uses hybrid retrieval and multi-objective optimization to overcome this challenge. **We hypothesize that our multimodal spatial, semantic, temporal, and visual integration will outperform text-only RAG for real-world POI recommendation.**

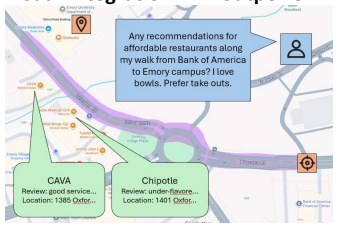

Figure 1: A real-world spatial reasoning question with nearby spatial objects.

Methods/Approach

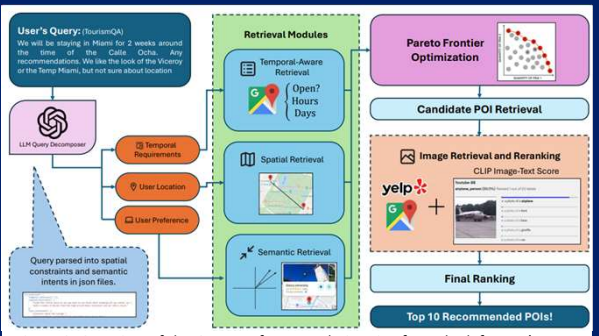

Figure 2: Overview of the GeoLens framework. Starting from the left, user's query was parsed into spatial constraints and semantic intent by a LLM decomposer. Our retrieval modules filter the POIs and was optimized using pareto frontier. The right shows the reranking process that utilizes images retrieval to improve the model's understanding of the POI. GeoLens ultimately generates the top 10 recommended POIs.

Results/Analysis

We compared GeoLens against the following established state-of-the-art baselines: **Spatial RAG (GPT4-Turbo), GeoLLM, naïve RAG**, and up-to-date LLMs including **GPT-5, Gemini-2.5, and Claude-4.5**. As shown in Table 1, GeoLens (bolded) achieves near **second-best performance** across almost all evaluation metrics on the TourismQA-Miami dataset, with particularly strong **ranking precision** and **answer completeness**. On TourismQA-NYC, it maintains competitive performance relative to mid-scale LLM baselines Across both datasets, **Spatial RAG (underlined) exhibits the strongest overall scores**.

Dataset	Method	Precision				Recall				F1				NDCG			
		@1	@3	@5	@10	@1	@3	@5	@10	@1	@3	@5	@10	@1	@3	@5	@10
TourismQA-Miami	GeoLens	**0.361**	**0.352**	**0.322**	**0.258**	0.041	**0.128**	**0.174**	**0.233**	0.071	**0.150**	**0.182**	**0.202**	**0.361**	**0.363**	**0.349**	0.334
	GeoLLM	0.316	0.272	0.237	0.205	0.037	0.095	0.125	0.182	0.059	0.109	0.131	0.162	0.316	0.320	0.327	0.360
	Spatial-Rag(GPT4-Turbo)	0.515	0.424	0.376	0.294	0.103	0.204	0.264	0.357	0.145	0.225	0.262	0.266	0.515	0.485	0.498	0.508
	Naive RAG	0.297	0.225	0.205	0.197	**0.081**	0.112	0.149	0.230	**0.101**	0.110	0.131	0.164	0.297	0.275	0.285	0.340
	GPT-5	0.152	0.152	0.164	0.221	0.002	0.007	0.013	0.035	0.005	0.014	0.024	0.059	0.152	0.169	0.227	**0.451**
	Gemini2.5	0.121	0.081	0.094	0.076	0.021	0.084	0.134	0.158	0.031	0.068	0.094	0.086	0.121	0.110	0.134	0.134
	Claude-4.5	0.237	0.149	0.153	0.147	0.004	0.007	0.012	0.023	0.007	0.013	0.022	0.040	0.237	0.208	0.271	0.407
TourismQA-NYC	GeoLens	0.335	0.294	0.295	0.306	0.017	0.044	0.076	0.154	0.032	0.071	0.109	0.180	0.333	0.306	0.311	0.337
	GeoLLM	0.365	0.329	0.302	0.273	0.017	0.043	0.061	0.103	0.031	0.069	0.091	0.133	0.365	0.363	0.371	0.425
	Spatial-Rag(GPT4-Turbo)	0.567	0.488	0.456	0.425	0.027	0.061	0.091	0.169	0.048	0.100	0.138	0.215	0.567	0.517	0.507	0.557
	Naive RAG	**0.492**	**0.445**	**0.425**	**0.418**	**0.021**	**0.056**	**0.089**	**0.168**	**0.040**	**0.093**	**0.135**	**0.217**	**0.492**	**0.455**	**0.453**	**0.506**
	GPT-5	0.441	0.417	0.410	0.382	0.019	**0.056**	**0.089**	0.163	0.035	0.091	0.133	0.202	0.441	0.413	0.388	0.304
	Gemini2.5	0.254	0.228	0.224	0.200	0.001	0.002	0.004	0.006	0.002	0.004	0.007	0.012	0.254	0.230	0.225	0.199
	Claude-4.5	0.257	0.241	0.237	0.215	0.012	0.029	0.046	0.065	0.022	0.048	0.071	0.107	0.257	0.300	0.364	0.499

Table 1: Performance comparison on TourismQA-Miami and TourismQA-NYC datasets.

Table 2 presents an evaluation of GeoLens on ground truths with or without image, and we observed that incorporating images into the evaluation logic consistently improved our results. This performance uplift further shows that **visual reranking module captures critical ambience information** that text descriptions alone may miss.

Ground Truth Setting	Precision ↑				Recall ↑				F1 ↑				NDCG ↑			
	@1	@3	@5	@10	@1	@3	@5	@10	@1	@3	@5	@10	@1	@3	@5	@10
Review Only	0.361	0.352	0.322	0.258	0.041	0.128	0.174	0.233	0.071	0.150	0.182	0.202	0.361	0.363	0.349	0.334
Review + Image	0.371	0.362	0.331	0.266	0.042	0.132	0.179	0.239	0.073	0.154	0.187	0.207	0.371	0.373	0.359	0.344

Table 2: Performance of GeoLens model on TourismQA-Miami dataset across 2 ground truth settings.

The impact of each retrieval method is illustrated in Figure 3 and 4, where the full GeoLens system achieves highest result compared to all other ablated variants. This observation confirms that **spatial, semantic, and temporal retrievals all contribute meaningfully to the recommendation quality.**

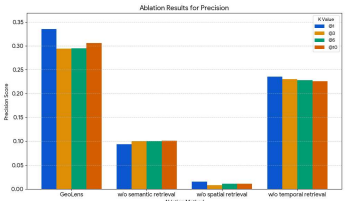

Figure 3: GeoLens Precision Ablation Bar Graph

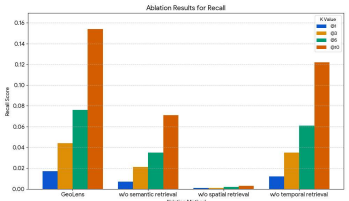

Figure 4: GeoLens Recall Ablation Bar Graph

Conclusions/Future Directions

Our hypothesis was supported as GeoLens demonstrated that integrating spatial, semantic, temporal, and visual retrievals within a single RAG framework leads to more reliable geospatial question answering. By incorporating Pareto frontier optimization and image reranking, GeoLens can balance multiple objectives and generate high quality POI recommendations.

Despite its effectiveness, GeoLens currently **relies on a static preprocessed POI database** and is **limited in handling fine-grained constraints** such as pricing and cuisine styles. **Ambiguous spatial requirements** in user queries, **modality-based evaluation bias**, and **dense semantic insufficiency** also remains a challenge for precise grounding as shown in the error distribution pie chart in Figure 5.

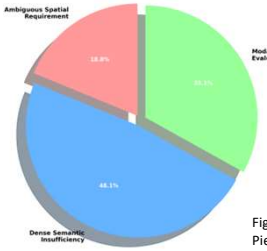

Figure 5: Error Distribution Pie Chart.

Future work should focus on:

- Scaling the evaluation pipeline to ensure stability and generalizability across larger metropolitan regions by **improving ground truth benchmarking on larger datasets**.
- Address data staleness, developing a dynamic POI would continuously update metadata and integrate real-time signals from external APIs into a **unified local datastore**, reducing reliance on static preprocessing and ensuring up-to-date recommendations.

Acknowledgements

Authors would like to thank Dr. Choi and Grace for detailed feedback and guidance on this research. We also gratefully acknowledge the authors of prior foundational works on RAG, geospatial reasoning, multimodal retrieval, and more that informed this study, including Spatial RAG, GeoLLM, OmniGeo, etc.