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

Mingyang Li*, **Henry Ning***, **Ethan Yang***
Department of Computer Science, Emory University
Atlanta, Georgia, USA
{mingyang.li, henry.ning, ethan.yang}@emory.edu

Abstract

Geospatial Question Answering (QA) and tourism recommendations are important due to their importance in location-based decision making and recommendation. Recent Retrieval-Augmented Generation (RAG) approaches have achieved remarkable results in geographic reasoning. However, current methods still struggle to handle multimodal signals while balancing explicit spatial and temporal constraints. This paper presents GeoLens, a multimodal RAG framework that uses hybrid retrieval and multi-objective optimization to overcome this challenge. GeoLens demonstrates superior results against recent Large Language Models (LLM), outperforming the best LLM baseline by 52.3% in Precision@1 on TourismQA-Miami, and competitive results against previous specialized system, exceeding GeoLLM by 49.5% in Recall@10 on TourismQA-NYC, with a light-weight LLM backbone. To the best of our knowledge, this is the first work to integrate multimodal signals into geospatial RAG. Our framework is useful for real-life geospatial recommendation applications with scalable cost.

1 Introduction

Geospatial reasoning is central to everyday applications such as city routing, travel planning, and location-based recommendations. Unlike abstract spatial reasoning problems studied in robotics or computer vision, geospatial reasoning requires interpreting large-scale, real-world data combined with rich semantics (Liu et al., 2025). For example, routing decisions must weigh not only the geometry of road networks, but also temporal constraints such as congestion patterns, whereas travel recommendations must balance spatial efficiency with subjective factors such as user reviews and textual descriptions. Existing GIS tools cannot process natural language while LLMs lack spatial grounding, and retrieval-augmented generation (RAG) appears to be the promising paradigm for bridging this divide by coupling LLMs with external data sources.

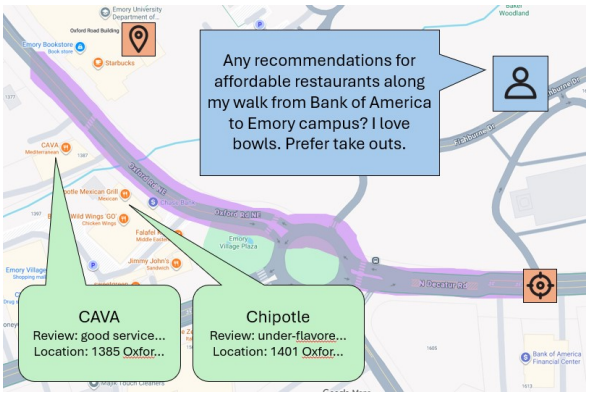

Figure 1: A real-world spatial reasoning question with nearby spatial objects. Areas that satisfy the spatial and semantic constraint are highlighted in purple.

Despite progress in retrieval-augmented generation (RAG), point-of-interest (POI)–oriented geospatial QA remains challenging. Natural language prompts often intertwine spatial constraints such as proximity, direction containment with semantic intent (e.g., cuisine, ambience), yet naïve RAG pipelines typically retrieve only text, overlooking the geometric precision of spatial computing and the contextual grounding available in imagery. This gap makes it difficult to resolve vague or descriptive queries such as "a restaurant overlooking the runway near the airport", where the system must jointly reason over spatial feasibility, semantic alignment, and visual cues (Luo et al., 2020; Staniek et al., 2024; Li et al., 2025; Qi et al., 2024; Deng et al., 2025).

Recent studies have begun to address this challenge from complementary angles. GA-LLM (Liu et al., 2025) demonstrated that injecting geographic coordinates and POI transition knowledge into LLM improves the next-POI recommendation, highlighting the benefit of explicit spatial embeddings. Building on this, Spatial-RAG(Yu et al.,

2025) extends spatial awareness to the larger real-world by combining a spatial retriever with semantic retrieval, enabling models to handle distance, direction, and containment constraints. Pushing further, OmniGeo(Yuan et al., 2025) introduces multimodal inputs, integrating text, spatial metadata, and imagery to support reasoning in both linguistic and visual-spatial cues. However, little research has explicitly integrated visual information into hybrid RAG strategies for geospatial QA.

In this paper, we propose GeoLens, a novel framework that utilizes hybrid retrieval strategy that unifies spatial, semantic, and image-based retrieval within a multi-stage pipeline. GeoLens explicitly addresses the multi-objective properties of real-world geospatial QA, ensuring that retrieved candidates not only satisfy geometric feasibility and semantic intent, but also leveraging imagery to enrich contextual understanding. Our contributions are summarized as follows:

1. Sparse and Dense semantic Retrieval: We enhance the semantic retrieval module of existing baselines, such as Spatial-RAG, by integrating sparse retrieval (BM25) alongside dense embeddings.

2. Temporal Scoring: We introduce a novel temporal scoring mechanism that filters candidates based on real-time operational status, opening hours, and day matching

3. Image Retrieval: GeoLens is the first geospatial recommendation system to integrate visual signals into the RAG pipeline.

We anticipate that GeoLens achieves higher retrieval accuracy in Geospatial QA tasks, compared to the retrieval approach without image understanding. By combining these modalities, GeoLens delivers more robust, trustworthy, and high-quality Point-of-interest recommendations in response to users' natural language prompt.

2 Related Work

2.1 Geospatial Recommendation

Geospatial question answering (QA) that joins spatial constraints and semantic preferences is still an emerging field. While many prior systems excel at either semantic-only RAG or spatial computing in GIS, relatively few unify different dimensions into a multi-objective pipeline. Recent work begins to close this gap with spatially grounded LLMs, spatial-RAG frameworks, and geospatial MLLMs. Overall, only a few works have tackled the full geospatial QA task, though several works have been done on similar tasks.

Yu et al. (2025) introduces Spatial-RAG, extending RAG to geospatial QA with a sparse–dense hybrid retriever and multi-objective generation that balances spatial and semantic relevance. Their design demonstrates strong retrieval, but offers limited handling of image information of POI and tool integration. Yuan et al. (2025) proposes OmniGeo, a multimodal LLM for a wide range of GeoAI tasks. While comprehensive in scope, it prioritizes benchmark demonstrations over task-specialized conversational retrieval with explicit spatial database operators. Liu et al. (2025) present GeoLLM, a vision-language recommendation model that links imagery with geographic concepts. This work advances image-grounded reasoning but does not address POI-centric, text+GIS QA.

Despite strong progress, challenges remain: (1) insufficient fusion of multi-modalities, spatial operators with semantic retrieval, (2) weak evaluation on real-world signals such as opening hours and travel times.

Our work offers complementary efforts in solving these issues because GeoLens treats geospatial QA as a multi-objective task. It parses natural language, performs hybrid candidate generation, applies multi-stage re-ranking, and supports tool-augmented grounding with APIs such as Google Maps, which enables richer handling of dialogue, geometry retrieval, and dynamic image sources compared to prior state-of-the-art approaches focused on static retrieval, or image-only grounding.

2.2 Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm that enhances the capabilities of large language models (LLMs) by incorporating external knowledge sources during the generation process (Lewis et al., 2021). For knowledge-intensive tasks such as question answering, RAG provides access to up-to-date, domain-specific, or spatially grounded information (Gao et al., 2024). A typical framework consists of retrieval, integration, and generation (Cheng et al., 2025); however, in geospatial QA, each stage faces unique challenges.

First, user queries intertwine spatial constraints (e.g., proximity, direction) with semantic preferences (e.g., cuisine type) (Yu et al., 2025). Second, retrieved candidates must be not only textually rel-

evant but also geometrically and visually grounded to support reasoning (Yuan et al., 2025). Third, integrating multimodal data—including text, spatial metadata, and imagery—requires complex fusion strategies to ensure coherence and accuracy (Zhao et al., 2023).

To address these, modern RAG systems adopt hybrid retrieval strategies (Yu et al., 2025), multi-stage re-ranking for spatial-semantic balance (Yoran et al., 2024), memory-augmented architectures for multi-stage coherence (Shinwari and Usama, 2025), and tool-augmented pipelines with live APIs for routing or business hours (Gao et al., 2024). More broadly, advances in adaptive retrieval and hallucination mitigation also inform robust design (Gao et al., 2024).

Our proposed framework, GeoLens, builds upon these advancements. It unifies spatial, semantic, temporal, and image-based retrieval in a multi-stage pipeline and explicitly models the multi-objective nature of geospatial QA. Retrieved candidates must simultaneously satisfy geometric, semantic, and temporal constraints while leveraging visual context to enhance trustworthiness. This represents a step toward robust, conversational, and multimodal RAG systems for real-world geospatial applications.

2.3 Data

1. **Text-framed Spatial Reasoning Benchmarks**: LLMSpatialBench (Xu et al., 2024) evaluates LLMs on text-only spatial reasoning (relative position, direction, adjacency, containment), revealing strengths in semantic understanding but no exposure to explicit map geometries. This type of issus is well tackled by Naive RAG, which grounds response in retrieved facts(Yu et al., 2025). 2. **Map-grounded Geospatial QA Benchmarks**: MapQA (Li et al., 2025) poses open-domain QA over map data with subsets for adjacency and proximity; instances bind NL queries to points, polylines, polygons and require geometric operators such as containment, distance thresholds. Text-to-OverpassQL (Staniek et al., 2024) provides a natural-language interface to OSM via OverpassQL programs that retrieve spatial entities under explicit topological/distance constraints, which is often used as a retrieval source when building map-QA evaluations. Given the fact that our model includes geospatial information retrieval through API, our project could facilitate the development in performance on this type of benchmark. 3.**Real-world Geospatial Evaluation** : TourismQA (NYC, Miami) from Spatial-RAG (Yu et al., 2025) collects noisy, user-generated POI questions from TripAdvisor, and travel forum, which mixes semantic preferences like category/reviews with spatial constraints (near/along a route/inside a region), which puts emphasis on constraint satisfaction and retrieval ranking in realistic tourism scenarios. Moreover, TravelPlanner (Xie et al., 2024) provides itinerary-style, map-grounded tasks with multi-step spatial constraints and goal satisfaction; while broader than pure QA, it serves as a strong real-world evaluation for route/POI feasibility and plan coherence. In this line of perspective, our model is tested on real-world evaluation dataset, which complements the discrepancy in evaluation within this field.

3 Method

3.1 Overview

As shown in Figure 2, given a natural language query q from TourismQA, GeoLens retrieves relevant Points-of-Interest (POIs) and produces truth-grounded response by integrating evidence from multiple modalities. The framework consists of four coordinated components— LLM query parsing, our retrieval methods, Pareto frontier optimization, and image retrieval. The pipeline begins with query parsing, where the system decomposes q into:

- Spatial constraints (e.g., “within five miles of downtown Atlanta”),
- Semantic constraints (e.g., “vegan cafés with outdoor seating”), and
- Temporal-aware constraints (e.g., “open after 10 PM on weekends”)

by a LLM decomposer. The parsed constraints guide the initial candidate generation: the spatial retriever filters geometrically feasible POIs, the semantic retriever identifies those that best align with the user’s intent based on dense text embeddings, and the temporal-aware retriever ensures only time-compatible candidates are considered. We used the Pareto frontier optimization strategy, which optimizes combination of spatial retrieval scoring and semantic retrieval scoring to make sure our candidate list is strong enough. Finally, in the rerank process, LLM leverages images retrieval to facilitate its contextual and ambience understanding of the POI to rerank the results. After

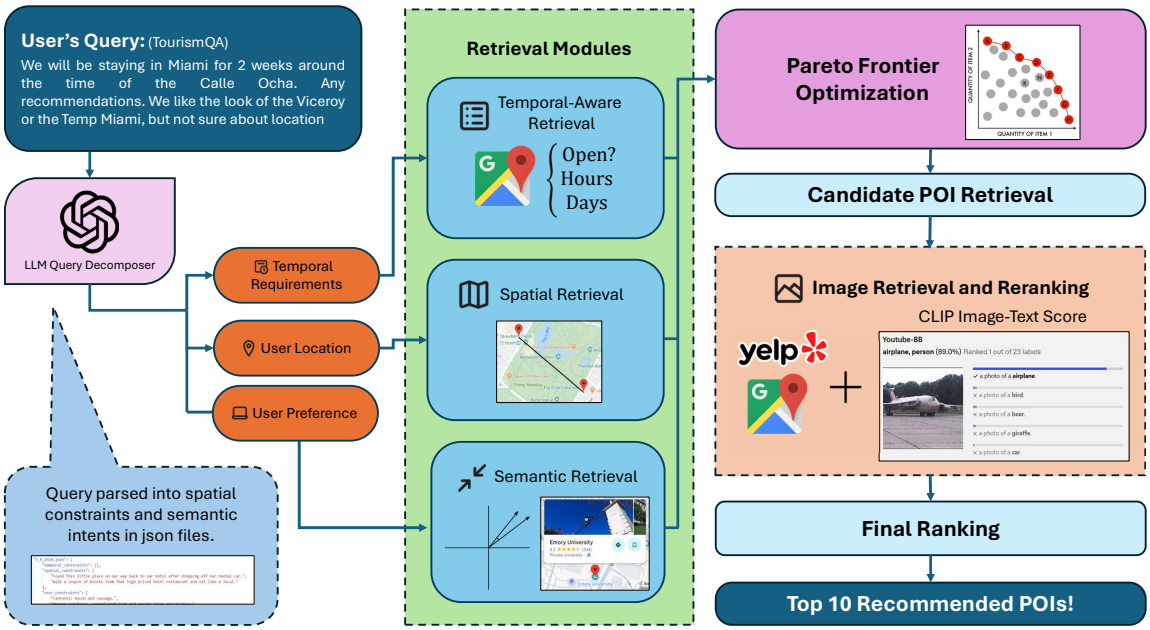

Figure 2: Overview of the GeoLens framework. Starting from the left, user's query was parsed into spatial constraints and semantic intent by a LLM decomposer. Our retrieval modules filter the POIs and was optimized using pareto frontier. The right shows the reranking process that utilizes images retrieval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

we have obtained 10 final POIs, we evaluate them using metrics that can evaluate recommendation qualities such as precision, recall, F1, and NDCG. Our model's performance is then used in comparison with current most cutting-edge LLM(GPT-5, Gemini-2.5, Claude-4.5) as well as Naive RAG and existing recommendation models(GeoLLM, Spatial-Rag). Ablation studies are also included in the experiment section to illustrate the effectiveness of each module within our system.

3.2 Spatial Retrieval

The spatial retrieval formula is designed to convert geographic distance into a normalized relevance score as shown below:

$$S_{\text{spatial}} = \frac{1}{1 + d(g_{\text{ref}}, g_{\text{cand}})}$$

S_{spatial} is the final spatial score and the distance function d calculates the shortest distance between two geometric objects, where g_{ref} is the reference geometry from the user's query and g_{cand} is the candidate geometry.

Since a shorter distance implies higher spatial relevance, a reciprocal form $1/(1 + d)$ is used to invert this relationship thus that POIs closer to the target area receive a score near 1, while distant ones receive scores near 0. The $+1$ in the denominator ensures stability in the algorithm by preventing division by 0 when the POI coincides exactly with the query center.

3.3 Dense Semantic Retrieval

The semantic relevance between a user's query and a POI description is determined and quantified by the cosine similarity of their vector embeddings. We used

The semantic score, $\mathbf{S}_{\text{semantic}}(\mathbf{q}, \mathbf{d})$, is calculated as:

$$\mathbf{S}_{\text{semantic}}(\mathbf{q}, \mathbf{d}) = \cos(\theta) = \frac{\mathbf{q} \cdot \mathbf{d}}{\|\mathbf{q}\| \|\mathbf{d}\|}$$

where $\mathbf{q}$ and $\mathbf{d}$ are the vector representations of the user's query and the POI's descriptive review, respectively. The resulting score is the dot product

of the two vectors divided by the product of their magnitudes.

Due to computational resources limit, we use Qwen-3 Embedding 0.6B to embed all queries and POI review. Then, similar meanings appear as vectors pointing in similar directions. Therefore, the cosine of the angle between them captures their relation independent of the text length. A value of 1 would indicate perfect semantic agreement, whereas a 0 would mean no correlation between the user's query and the POI's description.

3.4 Sparse Semantic Retrieval

Even though dense retrieval can effectively extract Places of Interest (POIs) with reviews similar to users' queries, users' long queries often introduce irrelevant noise. This makes dense retrieval susceptible to *semantic drift*, where the generated embeddings are diluted by non-essential terms, leading to the retrieval of semantically related but factually incorrect results (Thakur et al., 2021). To address this limitation, we employ sparse retrieval via the BM25 algorithm (Robertson et al., 2009) to complement the semantic search. By integrating exact keyword matching, BM25 acts as a lexical anchor, ensuring that specific user constraints (e.g., specific dish names or amenity types) are strictly adhered to—a hybrid strategy that has been shown to significantly outperform single-modality retrieval in noisy environments (Karpukhin et al., 2020; Lin et al., 2021).

$$\text{Score}_{\text{BM25}} = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + K} \tag{1}$$

$$\text{where } K = k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right) \tag{2}$$

3.5 Temporal-Aware Information Retrieval

To ensure factual temporal validity, GeoLens assesses whether each POI is active and open during the user's intended visit window. We introduce a novel temporal scoring, which relies purely on verified operational status, opening-hour, and opening-day alignment, which is obtained by Google API and Yelp API because google map api provides comprehensive information of POI's operational hours. Due to cost restraint, we have to use two API to get operational hours of all POIs given the large number of all POIs.

$$S_{\text{temporal}} = \frac{w_1 s_{\text{isOpen}} + w_2 s_{\text{hourMatch}} + w_3 s_{\text{dayMatch}}}{w_1 + w_2 + w_3} \tag{3}$$

where $O_{\text{status}} \in \{0, 1\}$ indicates if the POI is currently operational (e.g., `business_status = "OPERATIONAL"`), and $H_{\text{window}} \in \{0, 1\}$ verifies whether the POI is open at the user's query time or desired time window. The days match is also binary. If the query does not specify any temporal intent (e.g., "now", "open", "weekend"), the temporal weight λ_t is set to zero, and this module is skipped.

3.6 Pareto Optimizer

$$\mathbf{a} \prec \mathbf{b} \iff \begin{array}{ll} (1) & \forall i \in \{1, \ldots, k\}, f_i(\mathbf{a}) \leq f_i(\mathbf{b}) \\ (2) & \exists j \in \{1, \ldots, k\} : f_j(\mathbf{a}) < f_j(\mathbf{b}) \end{array}$$

The Pareto optimizer is designed to balance multiple retrieval objectives—spatial proximity, semantic relevance, and temporal validity—without requiring manual weighting. A candidate $\mathbf{a}$ is said to *Pareto-dominate* another candidate $\mathbf{b}$ if it performs no worse across all objectives and strictly better in at least one dimension, as defined above. In practice, GeoLens computes three normalized scores $(S_{\text{spatial}}, S_{\text{semantic}}, S_{\text{temporal}})$ for each candidate POI and identifies the non-dominated frontier of candidates that achieve optimal trade-offs among these metrics. The final Pareto-optimal set is then passed to the image reranking stage for fine-grained contextual filtering.

Compared to a simple weighted-sum fusion, Pareto optimization ensures fairness across heterogeneous retrieval signals, preventing any single modality (e.g., semantic similarity) from overwhelming others, which also enhances interpretability, as each retained candidate explicitly represents a balanced solution rather than a tuned scalar combination. The design generalizes to other multi-objective retrieval tasks where diverse criteria—such as accuracy, diversity, or freshness—must be optimized jointly without parameter-sensitive fusion.

3.7 Image Reranking

While the combination of spatial and semantic scoring effectively filters geometrically feasible and contextually relevant candidates, real-world geospatial QA often relies on subjective, non-textual cues that are best captured through imagery (e.g., assessing the "luxurious ambience" of a loca-

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

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**	**0.284**	**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

tion"). GeoLens incorporates an image-based retrieval in reranking stage to address this constraint. Images of POIs are web scraped from Yelp and Google Map. To ensure reliable visual-semantic alignment, images of 256 x 256 resolutions without blurring or exposure are kept.

For each candidate POI c_i generated by the hybrid spatial-semantic retrieval, we retrieve its primary associated image set $\mathcal{I}_i = \{I_{i,1}, I_{i,2}, \dots\}$. We employ a pretrained Vision-Language Model CLIP (Radford et al., 2021), to compute a visual-semantic alignment score $\mathcal{S}_{\text{image}}(q, I_{i,j})$ between the user's query q and each image $I_{i,j}$ since it allows for direct score computing between image and query. The overall image score for the POI is determined by the maximum alignment achieved:

$$S_{\text{image}}(q, c_i) = \max_{I_{i,j} \in \mathcal{I}_i} \mathcal{S}_{\text{VLM}}(q, I_{i,j})$$

where $\mathcal{S}_{\text{VLM}}$ is typically the cosine similarity of the text and image embeddings. This score captures the visual information of the candidate with respect to the descriptive elements in the query. Final candidates are reranked based on this image-alignment scoring produced by CLIP.

4 Experiments

4.1 Experiment Settings

Data Section: The experimental evaluation was conducted using two real-world tourism question-answering datasets: **TourismQA-NYC** and **TourismQA-Miami**(Yu et al., 2025). NYC has larger corpus, consisting of 9470 POIs and 17488 QA pairs while Miami has relatively smaller corpus, consisting of 2640 POIs and 133 QA Pairs.

Each entry consists of a user query paired with a set of ground-truth Points of Interest (POIs) that serve as the correct answers to that query. In terms of preprocessing, we followed same tradition. The questions are gathered from TripAdvisor posts and the reviews of restaurants, attractions, and hotels are sourced from travel forums and hotel booking platforms. POIs images are also scraped from TripAdvisor. Query vector database and POI reviews are embedded using open-source Qwen-3 Embedding-0.6B. Table 2 presents a summary of the corpus characteristics, notably the significant difference in size between the two regions.

Table 2: Overview of Datasets

Dataset	#POIs	#QA Pairs	#Images
TourismQA-NYC	9,470	17,448	18940
TourismQA-Miami	2,640	133	5280

Data Split: Just as all models under evaluation, including the proposed GeoLens framework, operate in a zero-shot or retrieval-augmented manner that does not require task-specific end-to-end training on the QA pairs themselves, the standard division into training, development, and test sets is not applicable for this evaluation. Thus, whole datasets (TourismQA-NYC and TourismQA-Miami) are utilized as the evaluation set to provide a comprehensive assessment of the models' ranking capabilities.

Evaluation Setting Two ground truth setting were used to evaluate our results. The first one is traditional way used in Spatial-RAG, where POI under queries are determined valid(1) or invalid(0) based on query, location, and review for locations. The other setting is where inject images of POI into input as well to evaluate valid or not. Based on two ground truth, we evaluate our results separately.

Hardware Configuration: All experiments were executed on an high-performance laptop integrated with an AMD Ryzen 7000-series processor

Table 3: Performance of the GeoLens model on the Tourism-Miami dataset. Two ground-truth settings are evaluated: review-only and review+image. (↑).

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

and an NVIDIA GeForce RTX 4060 Laptop GPU equipped with 16 GB of GDDR6 VRAM.

4.2 Evaluation Metrics

We evaluate the performance of all models using standard information retrieval metrics, focusing specifically on the ranking of relevant Points of Interest (POIs) retrieved for a given query. The results are reported at cutoff positions $k \in \{1, 3, 5, 10\}$.

Let R be the set of all relevant POIs for a query, and let L be the list of retrieved POIs, ranked from $i = 1$ to k. We define Rel_i as a binary indicator function where $\text{Rel}_i = 1$ if the POI at rank i is relevant, and 0 otherwise.

Precision Precision@k measures the fraction of retrieved results in the top k positions that are actually relevant. It is a critical metric for assessing the **quality** of the model's most immediate recommendations, ensuring high user satisfaction with the top results. However, precision only measures how accurate the top-k results are but ignores how many relevant POIs were missed. We complement this limitation by including recall and F1 to evaluate completeness.

$$P@k = \frac{1}{k} \sum_{i=1}^{k} \text{Rel}_i$$

Recall Recall@k measures the fraction of all relevant POIs in the dataset that are present within the top k positions of the retrieved list. This metric is essential for assessing the completeness of the retrieval process, ensuring the model does not miss a large portion of the potential correct answers. However, recall emphasizes coverage, which we compensate by reporting Precision@k and F1@k alongside recall to penalize systems that over-retrieve noisy POIs.

$$R@k = \frac{\sum_{i=1}^{k} \text{Rel}_i}{|R|}$$

F1-score The F1-score is the harmonic mean of Precision@k and Recall@k, providing a single, balanced measure of retrieval performance that accounts for both false positives and false negatives. Nonetheless, F1 balances precision and recall but remains insensitive to the ranking order of retrieved items, which is further complemented by NDCG@k in our setup that captures rank-sensitive performance.

$$F1@k = 2 \cdot \frac{P@k \cdot R@k}{P@k + R@k}$$

Normalized Discounted Cumulative Gain@k (NDCG) NDCG@k is a rank-aware metric that assigns higher scores to relevant POIs retrieved at earlier ranks. This is particularly important for a recommender task like tourism Q&A, as the highest-ranked suggestion must be the most appropriate answer. We assume binary relevance ($\text{Rel}_i \in \{0, 1\}$). NDCG depends on ranking quality and assumes perfect or binary relevance, making it unstable under noisy labels, which is balanced by F1 score to achieve better robustness.

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}$$

where:

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{\text{Rel}_i}{\log_2(i+1)}$$

IDCG@k is the ideal (maximum possible) DCG for the query, obtained by ranking all relevant POIs perfectly up to k.

4.3 Models

This section details the GeoLens and the baseline models used for performance comparison. The distinction between models highlights the contribution of incorporating explicit spatial and temporal awareness into the Retrieval-Augmented Generation (RAG) framework, directly addressing our research objective of grounding language models in structured geographic data.

GeoLens Configuration and Development: We evaluate our proposed framework, GeoLens, as described in Section 3. For the experimental

setup, the retrieval pipeline is configured to process the full candidate pool ($N = 2640$ for Miami, $N = 9470$ for NYC). Unlike stepwise filtering approaches, GeoLens first computes spatial, temporal, and semantic scores for all candidates. These scores are inputted into the Pareto Optimizer, which identifies the non-dominated frontier to select a balanced set of $K_{pareto} = 50$ candidates. These candidates are subsequently processed by the Image Reranker, which selects the final top $K = 10$ recommendations based on visual-semantic alignment.

Model Development and Modifications: During development, we observed that dense semantic embeddings were susceptible to irrelevant noise within complex user queries, leading to semantic drift. To address this, we incorporated sparse retrieval (BM25) to complement the dense retriever. This modification acts as a lexical anchor, ensuring that specific keyword constraints are strictly adhered to despite potential noise in the dense vector space.

Baselines and Comparisons: We compare GeoLens against the following established and state-of-the-art baselines. **Spatial-RAG (GPT4-Turbo)**: A strong RAG baseline that incorporates explicit spatial constraints during the retrieval phase. However, unlike GeoLens, it lacks the fine-grained temporal and multimodal image reranking capabilities. **GeoLLM**(Manvi et al., 2023): A specialized Large Language Model (LLM) fine-tuned for geospatial tasks. It relies purely on the model's internal knowledge and is used to assess the performance gain provided by external retrieval across all RAG-based methods. **Naive RAG:**(Lewis et al., 2020) This represents a standard RAG setup. It utilizes BGE-large-en-v1.5 embeddings and a standard cosine similarity search, serving as a lower bound for retrieval-augmented models. It deliberately lacks any explicit spatial or temporal awareness in its retrieval or ranking. **Up-to-Date LLMs**: GPT-5(OpenAI, 2025), Gemini-2.5(Comanici et al., 2025), Claude-4.5(Anthropic, 2025): These models were tested in a zero-shot, non-RAG setting using API. They establish the performance baseline of current state-of-the-art closed-source models when they do not have direct access to the structured POI knowledge base.

4.4 Performance Comparison between Models

Table 1 presents quantitative comparisons of GeoLens and several baseline models on the TourismQA Miami and NYC datasets. The table reports precision, recall, F1-score, and discounted accumulative gain (NDCG) cutoff positions @1, @3, @5, and @10. Baselines include GeoLLM, Spatial-RAG utilizing GPT4-Turbo, naive RAG, GPT-5, Gemini-2.5, and Claude-4.5.

On the TourismQA-Miami dataset, GeoLens nearly achieves the second-best results across almost every evaluation metric, effectively bridging the gap between the state-of-the-art Spatial-RAG and other baselines. For ranking precision, GeoLens attains a Precision@1 of 0.361, surpassing the third-best model, GeoLLM (0.316), by approximately 14.2%. Similarly, in terms of answer completeness, GeoLens achieves an F1@10 of 0.202, which represents a 23.1% improvement over the Naive RAG baseline (0.164). While Spatial-RAG (GPT-4-Turbo) remains the top performer (e.g., F1@5 = 0.262), GeoLens demonstrates robust generalization and competitive ranking quality (NDCG@10 = 0.334) without relying on a frontier-scale LLM backbone.

On the TourismQA-NYC dataset, GeoLens achieves Precision@1 = 0.335 and F1@10 = 0.180, maintaining close performance to other mid-scale LLM baselines such as Claude-4.5 (Precision@1 = 0.257, F1@10 = 0.107) and Gemini-2.5 (Precision@1 = 0.254, F1@10 = 0.012). Although Spatial-RAG (GPT-4-Turbo) exhibits the strongest overall scores (Precision@1 = 0.567, F1@10 = 0.215, NDCG@10 = 0.557), GeoLens remains competitive in both retrieval precision and ranking stability, achieving NDCG@10 = 0.337, which surpasses Gemini-2.5 and approaches the performance of Claude-4.5.

However, in comparison with previous state-of-the-art recommendation frameworks(Spatial-RAG) in the geospatial domain, GeoLens still falls short in overall quantitative performance. In particular, Spatial-RAG achieves the best results across all evaluation metrics on both TourismQA-NYC and TourismQA-Miami, outperforming GeoLens in Precision, Recall, F1, and NDCG at every cutoff position.

Table 3 presents a comparative evaluation of GeoLens under two ground-truth settings: "Review Only" and "Review + Image". We observed that incorporating images into the evaluation logic consistently improved our results. Specifically, on the Miami dataset, Precision@1 increased from 0.361 to 0.371, and NDCG@10 rose from 0.334 to 0.344. This performance uplift confirms that the visual

Table 4: Ablation Results

Method	Precision				Recall				F1				NDCG			
	@1	@3	@5	@10	@1	@3	@5	@10	@1	@3	@5	@10	@1	@3	@5	@10
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**	**0.33 3**	**0.306**	**0.311**	**0.337**
w/o semantic retrieval	0.094	0.100	0.100	0.101	0.007	0.021	0.035	0.071	0.011	0.028	0.041	0.065	0.094	0.099	0.101	0.110
w/o spatial retrieval	0.015	0.008	0.011	0.011	0.001	0.001	0.002	0.003	0.001	0.001	0.003	0.005	0.015	0.001	0.011	0.011
w/o temporal retrieval	0.235	0.230	0.228	0.226	0.012	0.035	0.061	0.122	0.022	0.056	0.086	0.136	0.235	0.234	0.236	0.249

reranking module captures critical "ambience" information that text descriptions alone may miss. Consequently, candidates that satisfy user intent through visual cues (e.g., a specific dining atmosphere) are correctly prioritized, thereby enhancing the overall recommendation quality compared to text-only approaches.

5 Analysis

5.1 Performance Analysis

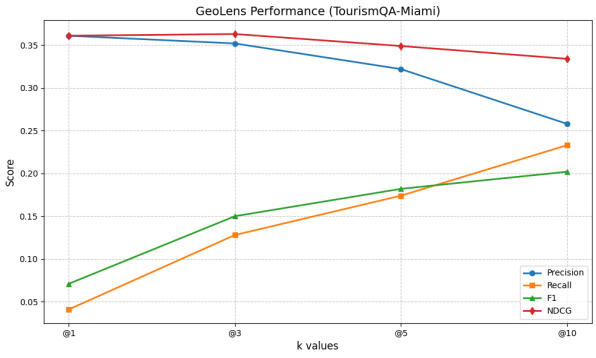

Figure 3: GeoLens Performance on Miami Dataset

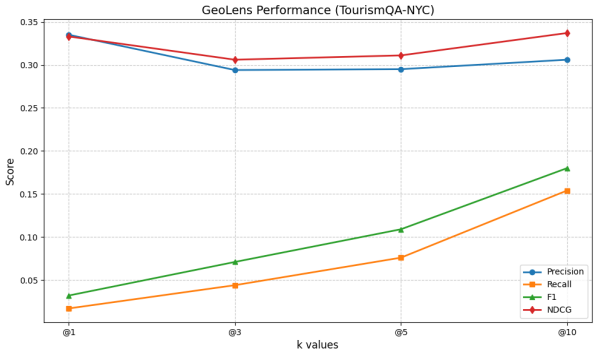

Figure 4: GeoLens Performance on NYC Dataset

Figures 3 and 4 illustrate the performance stability of GeoLens across varying retrieval depths. As k increases from 1 to 10, the model maintains high top-tier relevance, particularly in the Miami dataset where Precision starts strong at 0.361 at k=1 and NDCG remains robust, finishing at 0.334

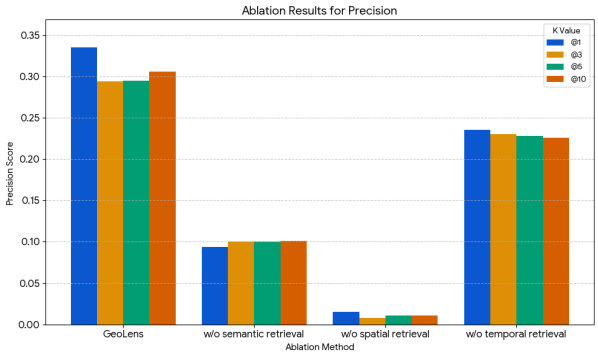

Figure 5: GeoLens Precision Ablation Bar Graph

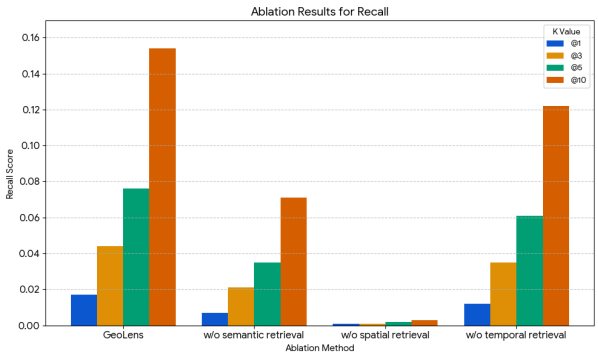

Figure 6: GeoLens Recall Ablation Bar Graph

at k=10. Similarly, for NYC, NDCG shows resilience, recovering to 0.337 at k=10, indicating that relevant POIs are consistently ranked near the top even as the candidate pool expands. In parallel, Recall exhibits steady growth, rising from 0.041 to 0.233 in Miami and 0.017 to 0.154 in NYC. This trajectory drives the F1-score upward, peaking at 0.202 for Miami, which confirms that the pipeline effectively captures a broader set of ground-truth POIs without sacrificing the ranking quality of the top recommendations

The ablation results from table 4 further emphasize the complementary effect of each retrieval module. According to Figure 5 and Figure 6, we observed that the GeoLens achieves highest result compared with all other ablated variants. Under precision metric, GeoLens gets a evenly distributed precision score across various **pass@k**. Under the

Recall metric, GeoLens improves recall rate exponentially as the **pass@k** increases.

The results clearly identify the following key findings regarding the GeoLens framework. First, GeoLens achieves performance metrics that are comparable to the frontier-scale LLMs on both TourismQA-Miami and TourismQA-NYC. This demonstrates that hybrid retrieval and re-ranking architecture of GeoLens is highly effective as the framework's ability to utilize spatial, semantic, and imagery data for retrieval. The ablation studies confirm the complementary effect of the GeoLens modules. This improvement is primarily driven by the interaction between spatial and semantic constraints. Spatial retrieval restricts the search space to plausible geographic regions, while semantic retrieval differentiates similar venues. Without spatial grounding, the system retrieves semantically similar but geographically invalid POIs. On the other hand, without semantic grounding, the model retrieves nearby but thematically irrevelant locations. As a result, the full GeoLens model outperforms all ablated variants, which validates our hypothesis that combining multiple retrieval modalities is crucial for complex geospatial QA. The **w/o temporal** variant achieves that second-best results. Temporal retrieval improves the overall metrics because most TourismQA queries do not explicitly reference the POIs' opening hours. When temporal constraints exists, models without temporal scoring would frequently include closed POIs in our top 10 rankings. Thus, temporal retrieval shows high impact on these cases and reduces critical recommendation errors. This finding signifies the importance of all spatial, semantic, and temporal retrievals and image re-ranking in our current framework.

Another key implication of our results is that pareto frontier optimization does not simply balance modalities, but also prevent domination of any single retrieval module. For example, semantic intents alone would overweight popular or descriptively rich POIs, whereas spatial scoring along would favor proximity without considering relevance. Pareto frontier optimization ensures that a POI is selected only if it is strong across all modalities, enforcing multimodal fairness in retrievals.

Based on the results presented above, GeoLens's key strengths are its architectural efficiency. It achieves competitive results without relying solely on LLM's scale by using a sophisticated RAG design. GeoLens's utilizes LLM as query decomposer to extract spatial and temporal constraints, and filter all candidates POIs before re-ranking them. This significantly improves efficiency. In addition, semantic and imagery re-ranking enables the accurate semantic meaning of retrieved POIs.

5.2 Error Analysis

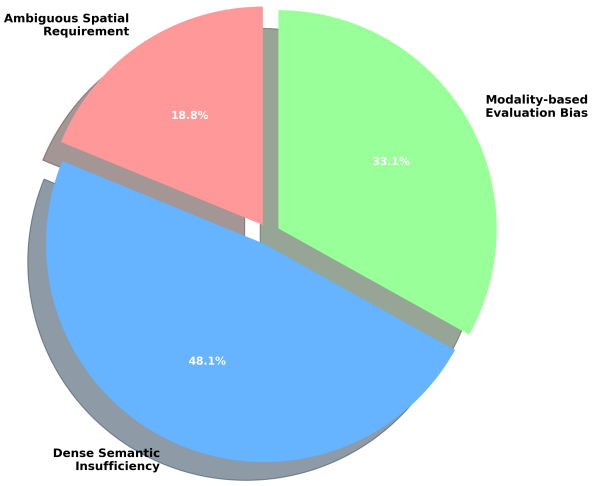

Figure 7: Error Distribution Pie Chart.

We categorize the observed errors into three distinct categories: Ambiguous spatial requirement, Dense Semantic Insufficiency, and modality-based evaluation bias. In total, there are 108 queries out of 438 queries that have POIs that fall into these three categories.

1. **Ambiguous Spatial Requirement:** 18.8% of misclassified predictions fall into this category. LLM decomposer fails to extract out reference spatial location from users' queries such as one Miami query: "I have been to a restaurant near the airport that overlooks a runway, but I can't remember the name of it. Can anyone help or make a suggestion for something else very close by." It results in [] outputed out by decomposer and spatial score returns 0 for all POIs. As a result, many POIs returned from database are not from Miami.

   Additional examples demonstrate this issue in NYC queries as well. In query NYU-125, the user mentions "taking LIRR into city from home" where the decomposer extracts only the ambiguous term "city" as the spatial reference. The geocoding API fails with "Could not find coordinates for 'city'", resulting in spatial scoring of 0 POIs and default

spatial scores of 0 for all results. Similarly, in query NYU-55, the decomposer extracts an unclear abbreviation "TOTR" (likely referring to "Top of the Rock"), which also fails geocoding. These ambiguous spatial references cause the system to lose geospatial precision. Even though our system falls back to the city(NYC/Miami) if no reference location is extracted, it still loses quality and precision in final POI recommendations.

2. **Dense Semantic Insufficiency:** 48.1%, the highest percentage of all three categories, of misclassified predictions are categorized as dense semantic insufficiency. This might be due to dense embeddings overweight broad similarities and overlooked more sparse and detailed attributes. We observed that dense semantic embeddings are not enough to retrieve most relevant POIs given that some POI recommendation has semantic score of 0.3/0.4, but spatial information is available. For instance, in query Miami-36, the user asks about transportation options: "I am taking a cruise out of Miami, but I am flying into and out of West Palm Beach. Should I rent a car to drive to and from Miami or is there another mode of transportation I can use to get to a Miami hotel near the port?" Despite successful spatial scoring (all POIs have spatial score between 0.9 and 1.0), the semantic scores remain low (ranging from 0.286 to 0.442), and the top results include irrelevant POIs such as "Burleigh Point Holiday Apartments" in Miami, FL, and other apartments, which are clearly mismatched to the query intent about Miami cruise port transportation.

   Another example is query NYU-55, where the user specifically requests "a good prix fixe menu with a couple of choices at lunch. American or French brasserie type restaurant with a budget of around $85pp." While the spatial requirement failed (geocoding "TOTR" failed), the semantic matching also performs poorly with scores between 0.293 and 0.454. The top results include "Rubens Empandadas Incorporated" (Argentinian, not French/American), "99 Cent Fresh Pizza" (fast food, not a brasserie), and "Hale and Hearty Soup" (casual soup place, not matching the upscale prix fixe requirement). These examples demonstrate that even when spatial constraints are satisfied, the dense semantic embeddings fail to capture nuanced query requirements such as cuisine type, price range, meal format (prix fixe), and dining style (brasserie), which suggests the need for stronger retrieval module.

3. **Modality-based evaluation bias** 33.3% of the misclassified predictions are due to modality-based evaluation bias. A significant portion of reported errors(0 for POI), though having scores for all retrieval modules(>0.7), stems from the *modality gap* between our model and the evaluator (GPT-4.1). We follow the evaluation tradition from Spatial-Rag(Yu et al., 2025) to determine whether the recommended POI is valid(1) or Invalid(0) based on the judgement of GPT-4.1 based on queries and review for POI. However, that way the images, a core process in our pipeline, is not included in the evaluation process. if the textual review for that POI does not explicitly mention "lighting," the text-only GPT-4 evaluator will be very likely to mark the recommendation as irrelevant (0).

5.3 Discussion

The experimental results highlight both the promise and current constraints of the GeoLens framework. Although GeoLens effectively integrates spatial, semantic, and visual modalities to support geospatial QA and outperforms recent large language model baselines, several limitations become apparent as GeoLens handles diverse query types and datasets.

Even though our system incorporates images to provide ambience information for Place-of-Interest recommendations, we preprocess these images before system execution as well as review for POI. It causes over-reliance on database issue when POI in a certain has limited review and POI. Our system falls short. Scalability wise, our system can only support real-life recommendations now but cannot deal with other geospatial questions answering.

Several promising extensions can further enhance the effectiveness and scalability of GeoLens. First, improving the ground-truth benchmarking process is critical. In this study, multimodal evaluation was primarily tested on the Miami dataset due to the huge computational costs of rebuilding the ground truth for the larger NYC dataset ($40,000+$ POIs). Future work should focus on scaling this evaluation pipeline to ensure stability and generalizability across larger metropolitan regions. Second,

to 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.

6 Conclusion

In this paper, we present GeoLens, a multimodal geospatial RAG framework that integrates spatial constraints, semantic intents, temporal availability, and visual information to improve real-life geospatial question answering. Our approach differs from traditional text-only RAG by incorporating images, enabling our system to reason beyond textual cues. Our model allows geospatial QA systems to overcome existing challenges in multimodal grounding by treating image information as integrated retrieval signals rather than isolated individual features.

Despite the progress, our model is still constrained by its reliance on a static, preprocessed POI database. To mitigate this, we plan to build a dynamic POI ingestion pipeline that continuously updates metadata, integrates external API sources into a unified local datastore in the future.

Acknowledgments

Authors would like to thank Dr. Choi and Grace for detailed feedback and guidance on this paper.

References

Anthropic. 2025. Introducing claude sonnet 4.5.

Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, Daoyu Wang, and Enhong Chen. 2025. A survey on knowledge-oriented retrieval-augmented generation. *Preprint*, arXiv:2503.10677.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Jieren Deng, Zhizhang Hu, Ziyan He, Aleksandar Cvetkovic, Pak Kiu Chung, Dragomir Yankov, and Chiqun Zhang. 2025. Imaia: Interactive maps ai assistant for travel planning and geo-spatial intelligence. *Preprint*, arXiv:2507.06993.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Zekun Li, Malcolm Grossman, Eric Qasemi, Mihir Kulkarni, Muhao Chen, and Yao-Yi Chiang. 2025. Mapqa: Open-domain geospatial question answering on map data. *Preprint*, arXiv:2503.07871.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2356–2362.

Zhao Liu, Wei Liu, Huajie Zhu, Jianxing Yu, Jian Yin, Wang-Chien Lee, and Shun Wang. 2025. Geography-aware large language models for next poi recommendation. *arXiv preprint arXiv:2505.13526*.

Hui Luo, Jingbo Zhou, Zhifeng Bao, Shuangli Li, J. Shane Culpepper, Haochao Ying, Hao Liu, and Hui Xiong. 2020. Spatial object recommendation with hints: When spatial granularity matters. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*.

OpenAI. 2025. Gpt-5 and the new era of work.

Feng Qi, Mian Dai, Zixian Zheng, and Chao Wang. 2024. Geodecoder: Empowering multimodal map understanding. *Preprint*, arXiv:2401.15118.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Haseeb Ullah Khan Shinwari and Muhammad Usama. 2025. Memory-augmented architecture for long-term context handling in large language models. *Preprint*, arXiv:2506.18271.

Michael Staniek, Raphael Schumann, Maike Züfle, and Stefan Riezler. 2024. Text-to-overpassql: A natural language interface for complex geodata querying of openstreetmap. *Transactions of the Association for Computational Linguistics*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. 2024. Travelplanner: A benchmark for real-world planning with language agents. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Liuchang Xu and 1 others. 2024. Evaluating large language models on spatial reasoning tasks. *Preprint*, arXiv:2404.09848.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. *Preprint*, arXiv:2310.01558.

Dazhou Yu, Riyang Bao, Ruiyu Ning, Jinghong Peng, Gengchen Mai, and Liang Zhao. 2025. Spatial-rag: Spatial retrieval augmented generation for real-world geospatial reasoning questions. *Preprint*, arXiv:2502.18470.

Long Yuan, Fengran Mo, Kaiyu Huang, Wenjie Wang, Wangyuxuan Zhai, Xiaoyu Zhu, You Li, Jinan Xu, and Jian-Yun Nie. 2025. Omnigeo: Towards a multimodal large language models for geospatial artificial intelligence. *Preprint*, arXiv:2503.16326.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023. Retrieving multimodal information for augmented generation: A survey. *Preprint*, arXiv:2303.10868.

A Appendix

A.1 Prompt Details

Listing 1: Prompt Query Decomposition

```
SYSTEM_PROMPT = """Your task is to act as a query decomposer.
      From the user's query, extract spatial and temporal
      requirements.

Respond only with a JSON object with two keys: spatial and
      temporal.

spatial: {
  "reference_location_text": string or null,
  "latitude": float or null,
  "longitude": float or null
}

temporal: {
  "start": string or null (YYYY-MM-DD or Monday-Sunday),
  "end": string or null (YYYY-MM-DD or Monday-Sunday),
  "time_of_day": string or null (morning, afternoon, evening,
       night, null)
}

Only extract the primary reference location and temporal
      requirements. Use null for missing values."""

USER_PROMPT = "Query: {query}"
```

Listing 2: Prompt: Ground Truth - Review + POI Images

```
SYSTEM_PROMPT = """You are an expert recommendation system
      evaluator. Your task is to determine if a recommended
      Point of Interest (POI) is a valid recommendation for
      the user's query.

You will be provided with:
1. User Query: The user's requirements.
2. POI Name.
3. Review: details for POI.
4. Visual Information: POI images.

Evaluation Logic:
- Analyze if the POI satisfies the query based on user queries,
       the Review for POI ,and images.
- Return 1 (Valid) if the POI is a good fit.
- Return 0 (Invalid) if the POI is irrelevant.

Respond ONLY 1(valid) or 0(invalid)
USER_PROMPT = """Query: {user_query}
POI Data: {
  "{poi_id}": {
    "name": "{name}",
    "review": {review_list}
  }
}
[Attached Images for this poi: {poi_id}]"""
```

Listing 3: Prompt:Ground Truth - Only Review

```
SYSTEM_PROMPT = """You are an expert recommendation system
      evaluator. Your task is to determine if a recommended
      Point of Interest (POI) is a valid recommendation for
      the user's query.

You will be provided with:
1. User Query: The user's requirements.
2. POI Name.
3. Review: details for POI.

Evaluation Logic:
- Analyze if the POI satisfies the query based on user queries,
       the Review for POI ,and images.
- Return 1 (Valid) if the POI is a good fit.
- Return 0 (Invalid) if the POI is irrelevant.

Respond ONLY 1(valid) or 0(invalid)
USER_PROMPT = """Query: {user_query}
POI Data: {
  "{poi_id}": {
    "name": "{name}",
    "review": {review_list}
  }
}
"""
```

A.2 Additional Case Studies

Figure 8: NYC Result Sample

User Query: *"Going to be in Manhattan for a broadway show this weekend and am looking for a recomendation for a good restaurant in the theatre district that is reasonably priced (30to50 per person) , good food and not a tourist trap. . . . would love to avoid the hotel restaurants and experience a real NY experience. . . Does such a place exist? Someone suggested Carmine's but they are totally booked. . . (like all kinds of food) Any suggestions would be greatly appreciated - picking one out of the 5000+ listed is a daunting task. . ."*

In this query, GeoLens initially decomposes the query into spatial ('Manhattan'), temporal ('start': '2023-10-14', 'end': '2023-10-15'), and preference constraints ('description_location_text': 'theatre district', 'pros': 'good food, reasonably priced (30to50 per person), non-tourist trap', 'cons': 'hotel restaurants'). It then scores a candidate pool of 9470 points of interest (POIs) across four dimensions—spatial proximity, temporal validity, keyword relevance (BM25), and semantic similarity—without initially filtering any candidates. A Pareto Frontier Optimization is subsequently ap-

plied to select the top 50 candidates that optimize the trade-offs between these scores. Finally, these 50 candidates undergo a CLIP-based image reranking process with original pre-trained weights to ensure visual relevance, reducing the selection to the final top 10 recommendations.

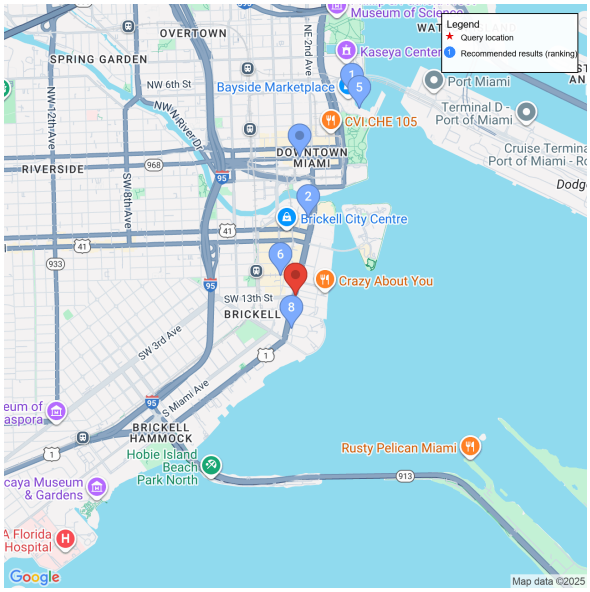

Figure 9: Miami Result Sample

User Query: *I have gone to the big Ft Lauderdale Winterfest Holiday Boat Parade many times, and love it. . . but have a friend staying near Biscayne Bay next weekend and noticed Miami has a boat parade. Since I live in Palm Beach County, I never knew about it; is it a big one (over 75 boats?) It looks like it is off Brickell Key, so should I assume it is an event only for those who have the wherewithal to live on that exclusive island (and pay exorbitant parking fees!) or does anyone have inside info/advice on any other good/safe place to see it from?*

In this query, GeoLens initially decomposes the query into spatial(Miami), temporal'start': None, 'end': None, 'timeofday': None, and preference constraints('description': 'viewing areas accessible to the public', 'pros': 'large number of boats expected (over 75)', 'cons': 'exclusive location, high parking fees). It then scores a candidate pool of 2640 points of interest (POIs) across four dimensions—spatial proximity, temporal(0 because no temporal constrains from user) validity, keyword relevance (BM25), and semantic similarity—without initially filtering any candidates. A Pareto Frontier Optimization is subsequently applied to select the top 50 candidates that optimize the trade-offs between these scores. Finally, these 50 candidates undergo a CLIP-based image reranking process to ensure visual relevance, reducing the selection to the final top 10 recommendations.

A.3 Details Regarding Retrieval Modules

Spatial Retrieval Configuration

We calculate the distance (d) in kilometers between the user's reference coordinates and the candidate POI. The relevance score follows the reciprocal decay function $S_{spatial} = 1/(1+d)$. In cases where the query decomposer returns a null reference location, the system defaults the coordinate inputs to the centroid of the target dataset (Miami or NYC).

Temporal-Aware Scoring Logic

The temporal score uses three weights: operational status (W_1), hour matching (W_2), and day matching (W_3). These weights are dynamically assigned such that $\sum W_i = 1$:

- **Full Constraints:** If status, hour, and day are specified, $W_1 = W_2 = W_3 = 1/3$.
- **Partial Constraints:** If a specific constraint is missing, its weight is set to 0. The remaining weights are normalized to sum to 1 (e.g., if only day is specified, $W_3 = 1$; if status and hour are specified, $W_1 = W_2 = 0.5$).
- **No Constraints:** If no temporal intent is detected, the module is bypassed.

Semantic Retrieval Parameters

The hybrid semantic pipeline utilizes the following configurations:

- **Dense Retrieval:** Uses the `Qwen-3-Embedding-0.6B` model for vector generation.
- **Sparse Retrieval:** Uses the BM25 algorithm with standard hyperparameters set to $k_1 = 1.5$ and $b = 0.75$. Additionally, we apply a weighted field matching scheme to align user preferences with POI attributes:
  - *User Pros* $\rightarrow$ *POI Pros*: $1.5\times$ weight
  - *User Description* $\rightarrow$ *POI Verdict*: $1.0\times$ weight
  - *User Negatives* $\rightarrow$ *POI Cons*: $-2.0\times$ penalty

Image Reranking Specifications

We utilize the pre-trained CLIP model (`ViT-B/32`)

initialized with original OpenAI weights. All input images are of resolution of 256×256 pixels to cater to the model. The final visual score is the cosine similarity between the query embedding and the image embedding calculated in real time.

A.4 Latency and Efficiency

In this paper, GeoLens processed 438 queries across the Miami (38 queries) and NYC (400 queries) datasets m. The system API cost includes GPT-4o-Turbo for query decomposition, Google Maps Geocoding API for location resolution, which leads to an average processing time of each query being approximately 15 seconds while per-query cost being approximately 0.0076$.