GEBENCH: BENCHMARKING TOOL-AUGMENTED AGENTS FOR GEO-ECONOMIC ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Geo-economic decision-making, like public-facility assessment and allocation, requires not only interpreting multimodal geospatial information but also actively discovering what data is missing and which geographic tools can supply it. While recent geospatial foundation models and multimodal large language models have improved perception and geo-grounding, the existing benchmarks predominantly focus on single-modal inputs and tasks related to remote sensing, leaving tool-driven economic and demographic knowledge discovery underexplored. This paper therefore introduces the Geo-Economic Benchmark (GEBench), a large-scale multimodal benchmark for predictive geo-economic reasoning grounded in remote sensing imagery, map-derived place context, and demographic priors. To the best of our knowledge, this is the first work to benchmark agentic, tool-interactive methods for solving geo-economic problems. GEBench formulates geo-economic tasks as question-answer pairs with multi-modal input and supports multi-round interaction. It comprises over 16,000 question-answer pairs across five building categories with their corresponding geo-economic prediction targets and spanning more than 6 countries. Building on GEBench, we develop an agentic reasoning pipeline in which an exploration agent plans feasible API queries and iteratively integrates results from mapping and demographic tools before downstream analysis and forecasting, enabling active information acquisition from Geo-Economic data. We systematically evaluate tasks including numerical prediction and localization, demonstrating that our agentic reasoning framework consistently outperforms direct model querying across diverse geo-economic scenarios.

1 INTRODUCTION

Geographic economic decision-making, such as location selection and capacity design of public facilities, is fundamentally a spatial reasoning problem over multi-modal information Zhang et al. (2024); Yuan et al. (2025). Effective assessment of rationality requires integrating complementary signals from multiple modalities, including satellite imagery Li et al. (2019); Yeh et al. (2020); Zhang et al. (2024), street-level observations Dai et al. (2023), mapped entities Li et al. (2022), and building attributes Yeh et al. (2020); Ahn et al. (2023b;a). In particular, geographical features inferred from remote sensing could be used to model and predict economic outcomes Weng et al. (2024). However, joint reasoning over such multimodal information remains challenging due to modality gaps, region diversity, and inconsistent data representations Ma et al. (2024); Wang et al. (2024); Sun et al. (2025). A practical geo-economic workflow also requires active knowledge discovery: given a question, the system must determine missing information (e.g., distribution of nearby buildings, map-scale context, population priors) and which geographic tools provide it, then query tools iteratively and integrate returned data.

Recent advances in geospatial foundation models and multimodal large language models have enabled spatial grounding, map-augmented reasoning, and tool-based geospatial analysis Dai et al. (2023). Agent-based frameworks further introduce capabilities for multi-step planning and coordinated reasoning across multimodal data sources Yang et al. (2025); Zhang et al. (2025a). Nevertheless, the existing systems are primarily evaluated on question answering, geo-grounding, navigation, or tool-execution tasks Sapkota et al. (2025); Ma et al. (2024); Huang et al. (2024); Zhao et al. (2025), rather than structured predictive settings that quantify geo-economic outcomes. As a result, current

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

evaluations underestimate the central capability required for geo-economic inference: tool-driven information acquisition and synthesis into calibrated numerical predictions.

To capture the interplay between geographic constraints and economic reward, we introduce the Geo-Economic Benchmark (GEBench), a comprehensive multimodal benchmark for predictive analysis of public-facility allocation. GEBench integrates satellite imagery, mapped entities, socioeconomic attributes, and population density estimates, enabling systematic evaluation of models that combine visual, geographic, and demographic information for structured numerical prediction. In addition to the dataset, we provide evaluation protocols and develop an agentic geo-economic reasoning workflow as a baseline framework for coordinated multi-source information reasoning. The proposed baseline method performs active information acquisition and leverages multiple collaborating agents to process multimodal inputs, ultimately producing structured geo-economic predictions and reports. By evaluating geo-economic prediction within a standardized benchmark, our work establishes a foundation for systematic comparison of multimodal geospatial reasoning approaches. To the best of our knowledge, this is the first work to focus on geo-economic problems through an agentic, tool-interactive perspective that emphasizes active information discovery with mapping and demographic tools. The contributions of this paper are summarized as follows:

- We formulate geo-economic assessment as a structured prediction problem that requires active knowledge discovery over mapping and demographic tools, followed by multimodal information aggregation into calibrated numeric outputs.
- We introduce GEBench, a comprehensive benchmark covering schools, hospitals, libraries, residential properties, and recreational trails, integrating satellite imagery, geographic context, and population density estimates with structured numerical targets across diverse interaction formats.
- We develop an agentic baseline that explicitly performs information planning and tool-based data acquisition, and establish a reproducible evaluation protocol for geo-economic reasoning.

2 Related Work

2.1 Geospatial Multimodal Language Models and Evaluation

Recent years have witnessed rapid progress in geospatial large language models and multimodal geospatial foundation models, accompanied by increasingly systematic evaluations of their capabilities Vatsavai (2024); Yang et al. (2025); Zhang et al. (2025a). Early efforts, such as GeoLM Li et al. (2023) and GeoLLM Manvi et al. (2023), explore grounding language models with map-derived context and geospatially aligned text, demonstrating the capacity of LLMs for understanding spatial semantics. Subsequent studies move beyond static question answering toward interactive and tool-augmented settings. For instance, GeoLLM-Engine Singh et al. (2024) provides a realistic environment abstraction for constructing geospatial copilots, which is capable of reasoning over tools and large-scale imagery. Parallel to these developments, a series of benchmarks, including Charting Roberts et al. (2024), GeoBenchX Krechetova & Kochedykov (2025), GeoAnalystBench Zhang et al. (2025b) evaluate geospatial reasoning and workflow capabilities, whereas GEOBench-VLM Danish et al. (2025), and SpatialLLM Abbas & Ahmad (2024) focus on multimodal spatial reasoning with vision-language as the inputs. Despite this progress, existing evaluations predominantly emphasize knowledge retrieval, geo-grounding, navigation, or tool execution, with limited attention to structured geo-economic prediction, which enables data-driven urban planning by converting multimodal geographic and socioeconomic features into calibrated, policy-relevant outcomes. Geo-economic prediction differs fundamentally from QA-style evaluation, which requires calibrated numeric outputs under information alignment.

2.2 Multimodal Agent Systems

LLM and VLM-based agents have evolved from single-model prompt programs to tool-augmented and multi-agent systems frameworks with planning, memory, and specialization Zhang et al. (2025a); Chi et al. (2025). In domain settings, agentic approaches have been explored in economics Horton (2023); Li et al. (2024) and urban/municipal applications Feng et al. (2025), including macroeconomic agents Li et al. (2024) and emerging perspectives on LLM-driven urban computing and planning workflows Pan et al. (2025); Li et al. (2025). Meanwhile, recent studies increasingly emphasize

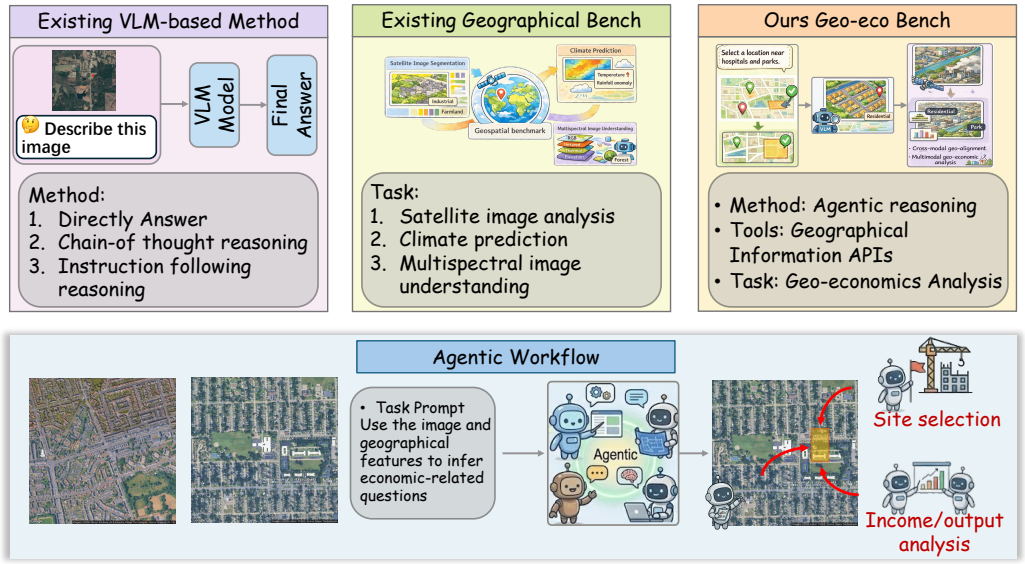


Figure 1: The proposed Geo-eco Benchmark (GEBench) aims to facilitate research on agentic reasoning for geo-economic analysis. As a first step, we implement a baseline agentic workflow and report its evaluation results on the benchmark.

agent-based orchestration over multimodal tools and data sources to support complex analytical workflows Zhang et al. (2025d); Singh et al. (2024). Agent systems such as MMSearch Jiang et al. (2024) and MindSearch Chen et al. (2025) adopt structured pipeline designs to enable access to external knowledge, thereby equipping LLMs with multimodal search capabilities. DeepEyes Zheng et al. (2026), Ground-R1 Cao et al. (2026), and Active-O3 Zhu et al. (2025) introduce tool-use-oriented data selection strategies and reinforcement learning to support LLMs with the ability of "thinking with images". REVPT Zhou et al. (2025) enhances multimodal LLMs with visual perception and reasoning by training them to dynamically leverage specialized visual tools. However, existing multimodal agentic systems primarily target retrieval, perception, or workflow execution, and lack support for structured geo-economic reasoning with numeric targets and multimodal information.

2.3 Geospatial and Geo-economic Database

Recent advances in geospatial platforms and datasets have enabled large-scale access to geographic information for mapping and spatial analysis. Platforms such as the Google Maps Platform provide APIs for retrieving satellite imagery and point-of-interest (POI) information, facilitating scalable geospatial data acquisition Google (b). OpenStreetMap (OSM) Haklay & Weber (2008) provides a large-scale, open, and crowd-sourced geographic database containing global mapping information. Commercial platforms further offer rich contextual services, such as property attributes and routing information Zillow Group; Mapbox; however, proprietary access and licensing policies may restrict reproducible research. Beyond geographic databases, several benchmark datasets have been introduced for learning visual and semantic attributes from urban environments, including city-scale building attributes Zhang et al. (2025c) and street-level imagery Cordts et al. (2016).

In addition to geospatial databases, several economic databases have been developed for research on social and economic problems. Our work leverages these valuable data sources for benchmarking the geo-economic ability of agentic frameworks. For example, WorldPop Lloyd et al. (2017) records the world population and density metrics, which are important in economic-related research. The United Kingdom government reveals registered property transactions with prices UK Government monthly. Apart from real-world databases, Li et al. (2024) introduce a simulation framework that can generate economic activity data from simulation.

Despite the diversity of existing resources, they are organized in a single-source manner, and our GEBench organizes these raw data into a dataset with integrated features across multiple sources

and builds a question-answer benchmark that aligns imagery, map context, and socioeconomic data. Fig. 1 illustrates that our GEBench integrates geospatial benchmarks and VLM methods to support structured geo-economic analysis, enabling evaluation of agentic reasoning for economic tasks.

3 GEBENCH

3.1 GEO-ECONOMICS TASKS

GEBench is designed to evaluate agentic geo-economic reasoning with active knowledge discovery. Given a geo-economic question, an agent should determine the necessary and feasible geographic, demographic and economic information and interact with mapping tools or existing databases to acquire this information, then reason with economic knowledge to estimate answers to these questions. Common examples of these questions include inferring the expected income, output, number of employees and visitors of public facilities such as hospitals, schools and parks, or comparing and selecting the best location for new buildings. Our benchmark includes geo-economic questions and ground truth answers for agentic reasoning setting, which covers both single-round and multi-round dialogue, including direct numeric prediction, facility localization in city-level remote sensing imagery, residential property price valuation, iterative reasoning with progressively revealed features, and temporal trend forecasting for geo-economic indicators such as annual visitors or enrollment.

3.2 DATASET OVERVIEW

Table 1: Building categories and the final geo-economic prediction targets of our dataset. By leveraging the ability of agentic reasoning, we can achieve the prediction of geo-economic indicators such as the number of students/teachers, revenue margin and net income, annual visitors, market prices, and daily visitor counts.

Category	*schools*	*hospitals*	*libraries*	*houses*	*park trails*
Targets	number of students/teachers	revenue margin and net income	annual visitors	market prices	daily visitor

Table 2: Comparison of GEBench with representative geospatial benchmarks and general agent evaluations across key dimensions.

Benchmark	**Geographical data**	**Multi-Round**	**Tool-Use**	**Agentic Reasoning**
Charting Roberts et al. (2024)	✓	-	-	-
GeoLLM-Engine Singh et al. (2024)	✓	✓	✓	-
GeoAnalystBench Zhang et al. (2025b)	✓	-	✓	-
GeoBenchX Krechetova & Kochedykov (2025)	✓	✓	✓	-
GEOBench-VLM Danish et al. (2025)	✓	-	-	-
VisualWebArena Koh et al. (2024)	-	✓	✓	-
OSWorld Xie et al. (2024)	-	✓	✓	-
GEBench (ours)	✓	✓	✓	✓

GEBench contains 16,000 dialogue records across five categories with category-specific geo-economic targets. The dialogues are generated from 5,380 eligible school locations, 999 hospitals, 18 libraries, 1,000 residential properties, and 39 trail sensors after filtering for usable geographic context. These source locations cover the United States, the United Kingdom, and Australia; because multiple templates may reuse one location, we report source-location counts separately from dialogue counts. Table 1 summarizes the categories and prediction targets in our dataset. The scale, diversity, and coverage of GEBench enable a systematic evaluation of the generalization and robustness abilities of models in dealing with geo-economic tasks.

Each record is organized as a structured dialogue with questions and answers, ranging from one to three rounds. This dialogue-based formulation of the dataset enables evaluation of both direct numeric prediction and iterative reasoning under progressively revealed geographic context, reflecting realistic geo-economic analysis workflows. Each record integrates multimodal geographic information sources, including target-level remote sensing imagery, city-level image, and structured geographic context

features such as the number of nearby point-of-interests and population-related indicators. These data provide complementary spatial, visual, and socioeconomic features for geo-economic reasoning.

Table 2 compares GEBench with representative geospatial benchmarks and general agent evaluations. Existing geospatial benchmarks mainly focus on perception-oriented tasks, such as satellite image understanding, object detection, and building-level semantic analysis. In contrast, GEBench targets geo-economic reasoning by integrating multimodal geographical evidence, socioeconomic information, and tool-assisted multi-round inference, enabling systematic evaluation of agentic reasoning for real-world economic analysis.

3.3 Evaluation Protocol and Metrics

Evaluation Protocol. The initial input to the agentic system contains only the geo-economic question. The Exploration Agent selects an available tool and requests the relevant cached information from the dataset. The tool outputs are pre-collected and stored in the dataset, but they are returned only after the corresponding tool is requested. The agentic system then reasons over the returned information and generates the final prediction.

Each question has a predefined output format, such as a numeric value, a selection result, or pixel coordinates. In the controlled comparison in Table 3, the direct-query baseline and the agentic pipeline receive the same images, POI counts, population values, and localization information. The direct baseline receives this evidence directly, while the agentic pipeline must request it through the interaction protocol. We record the intermediate outputs and tool traces for analysis.

Numeric Prediction. For school, library, residential property, and recreational trail categories, we evaluate numeric predictions using Mean Absolute Percentage Error (MAPE). For hospital financial indicators, we report Mean Absolute Error (MAE). Let y denote the ground-truth value and $\hat{y}$ the model prediction, MAPE is computed as: $\mathrm{MAPE}\% = \frac{|\hat{y}-y|}{|y|} \times 100$. Since the model might generate unstructured text, we additionally measure the fraction of items for which the required structured outputs can be successfully extracted using regex-based parsing.

Localization. For facility localization tasks, models predict pixel coordinates in city-level satellite imagery. We compute the Euclidean distance between predicted and ground-truth pixels: $d(\hat{\mathbf{p}}, \mathbf{p}) = \|\hat{\mathbf{p}} - \mathbf{p}\|_2$. Localization performance is evaluated using the average pixel distance and the percentage of predictions within a fixed threshold ($\tau = 100$ pixels).

Image Causal Removal. We additionally provide edited satellite images in which the target facility is marked by the instance masking and removed, as shown in Fig. 2. Given facility coordinates obtained from the Google Maps Platform API, the building instance is segmented using a SAM-based vision-language segmentation model Kirillov et al. (2023); Ravi et al. (2024). The resulting instance is then masked, and the occluded region is filled using DPM-Solver++ Lu et al. (2025) to produce a plausible, texture-consistent image. These edited images allow the benchmark to support counterfactual spatial reasoning and causal analysis, such as evaluating how geo-economic indicators might evolve under alternative land-use configurations, facility placement scenarios, or selecting target regions without direct visual features.

3.4 Benchmark Construction

The dataset is constructed through three stages: raw geo-economic data collection, multimodal dataset construction, and template-driven task generation.

Raw Information Sources. Across all categories, the benchmark integrates satellite imagery, socioeconomic features, and structured numeric targets from diverse public sources. Specifically, place-level and city-level map tiles are collected via the Google Maps Platform Static Maps API Google (b). Geographic context features, such as nearby point-of-interest counts and aggregated attributes, are obtained through the Google Maps Platform Places API Google (a). Population information is derived from the 2025 WorldPop gridded population dataset Lloyd et al. (2017) and is integrated across all facility locations.

Category-specific targets comprise student enrollment and teacher counts for schools, financial indicators from the AHRQ Hospital Financial Measures Database Agency for Healthcare Research

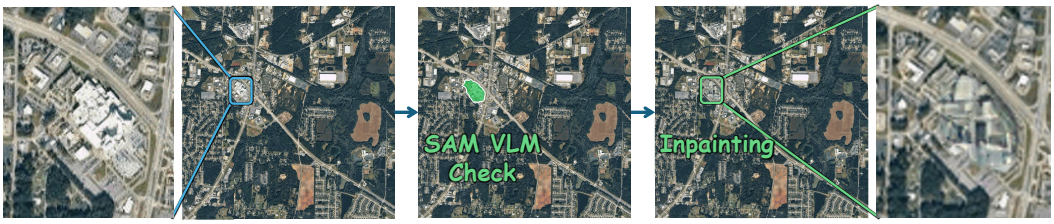


Figure 2: **Generation of instance-masked satellite imagery.** (a) Original satellite image containing the target facility. (b) Building instance segmented using a SAM-based VLM. (c) Masked region filled via texture-consistent inpainting.

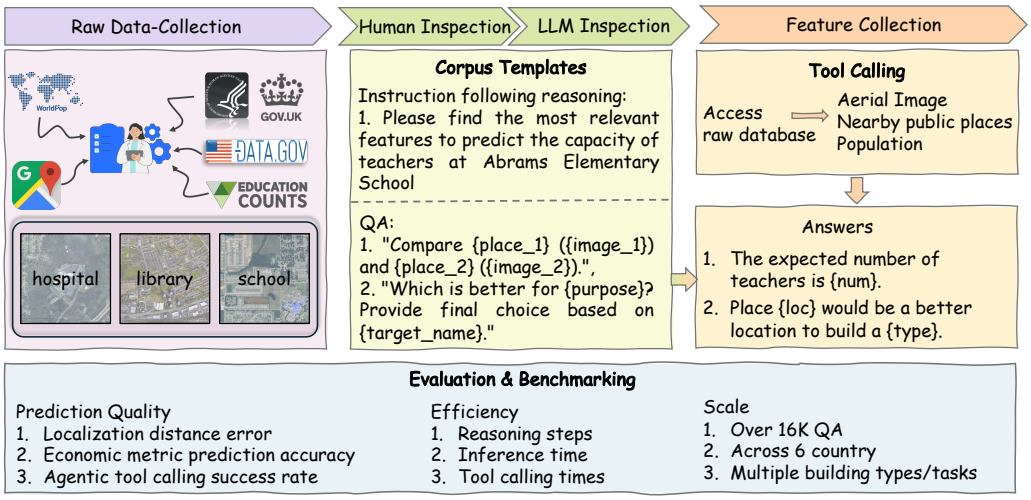


Figure 3: **The construction process of GEBench.** Raw geo-economic data are collected from online resources and transformed into geo-economic question-answer pairs via template-based generation.

and Quality for hospitals, annual visitor counts from public records for libraries, market price data for residential properties, and daily visitor count estimates for recreational trails.

Dataset Construction Pipeline. Starting from raw facility records with identifiers and geo-economic targets, we construct each benchmark item through a deterministic preprocessing pipeline. First, facility coordinates are validated and canonicalized through geocoding queries, converting facility names into standardized geographic coordinates and associated place metadata. Second, satellite imagery is retrieved via Static Maps to generate both place-level crops and city-level images for the localization task, while recording pixel-level ground-truth coordinates of facilities within city-scale images. Third, amenity and point-of-interest features are extracted via Places Nearby Search, aggregating nearby-place counts and summary statistics such as average ratings into structured context fields. Finally, we use pre-defined templates, including various question types to generate Question-answer pairs. During evaluation, the agentic system takes the question as input and the output is compared with the ground truth answer with various metrics based on the question type. Fig. 3 provides an overview of the dataset construction pipeline.

Template-Driven QA Generation. We construct benchmark dialogues using a template bank that defines diverse task formats and interaction styles. As illustrated in Fig. 4, each template integrates multi-scale satellite imagery together with structured geographic features to generate geo-economic questions, structured reasoning traces, and target answers. The benchmark supports four tasks. *Multi-round inference* progressively reveals contextual information and requires models to produce a final numeric prediction. *Localization* tasks require models to return pixel coordinates of the queried facility within city-level satellite images. *Direct lookup* involves single-turn queries for retrieving target attributes associated with a referenced location. *Property valuation* tasks estimate residential property prices based on visual building characteristics and surrounding demographic information.

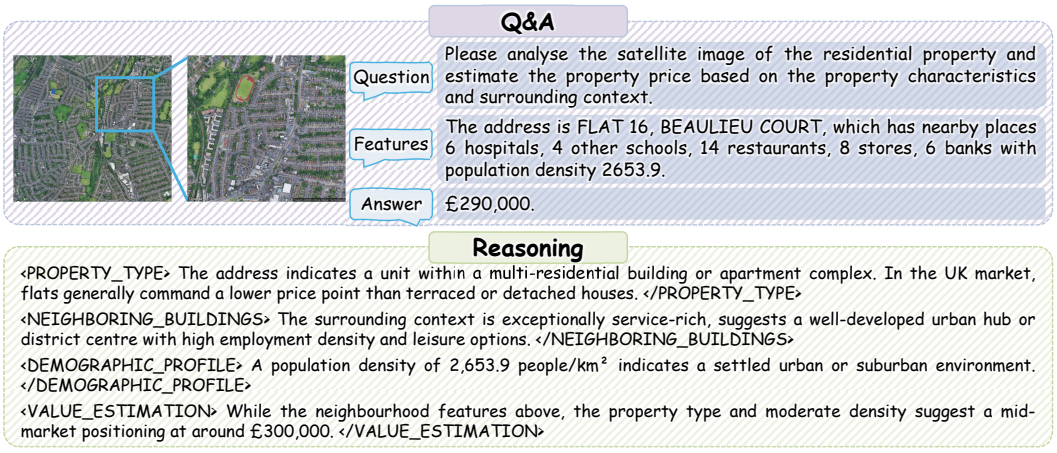


Figure 4: **Illustration of template-driven QA generation.** Multi-scale satellite images and structured geographic features are injected into question templates to produce geo-economic questions, structured reasoning steps, and final answers.

3.5 Baseline Framework for Agentic Geo-Economic Reasoning

We build a baseline agentic system for active geo-economic reasoning, including six agents for controlling the information flow and performing the reasoning to acquire final output, a set of pre-built tools for requesting the features, and an error check mechanism for the output. We will elaborate our agentic system design in the following.

Formal Task Statement. Each data item is an image-text question-answer pair $r = \{\mathcal{D}, \mathcal{I}, \mathbf{c}, y\}$, where $\mathcal{D} = \{(q_t, a_t)\}_{t=1}^{T}$ is a short dialogue, $\mathcal{I} = \{I_k\}_{k=1}^{K}$ are referenced images (place-level and/or city-level satellite patches), $\mathbf{c}$ is lightweight structured context, and y is the ground-truth numeric target (used only for evaluation). The model outputs scalar predictions $\hat{y}$ for schools, libraries, residential properties, and recreational trails, or tuples of financial indicators for hospitals, and pixel coordinates $\hat{\mathbf{p}} = (\hat{x}, \hat{y})$ for localization tasks.

Agent Reasoning Workflow Design. We implement each specialized agent with a fixed input/output message interface, while sharing a common VLM backbone across all agents. The workflow of the proposed Agentic Geo-Economic Reasoning framework is illustrated in Fig. 5. Concretely, the framework consists of six specialized agents: (1) Explore parses the initial intent and selects appropriate tools through predefined interfaces; (2) Satellite analyzes requested satellite imagery; (3) Context interprets POI, population, and map-based evidence; (4) Forecast performs numerical synthesis and estimation; (5) Evaluate computes task-specific metrics; and (6) Report generates the final structured response and reasoning trace. Each agent returns a structured message object with a predefined set of required fields, making intermediate states explicitly inspectable and improving robustness against error propagation from unconstrained free-form generation.

Active Information Collection and Agent Pipeline. Explore agent is responsible for active information acquisition. Instead of passively consuming fixed inputs, it iteratively analyzes the question intent and current information, decides what feasible map-derived signals are missing, queries external APIs accordingly, such as geocoding and nearby-place inventories from Places, additional tiles from Static Maps, and population priors from WorldPop, and reasons over returned results to update its internal information handling.

A Satellite Agent processes input image $\mathcal{I}$ and produces an analysis message $\mathbf{v}$, combining facility-scale features such as site size, surrounding land use, and connectivity with town-scale context. When information planning indicates that a broader context is needed, it can retrieve additional town-level map tiles from the Maps API and produce auxiliary visual artifacts such as a town map view and a building-distribution summary to support prediction. Context $\mathbf{c}$ is taken as prompt of the dialogue and produces an analysis $\mathbf{u}$ that maps local amenities/POI density to demand-related signals.

External Tools and Data Usage. During agentic reasoning, the framework incorporates tool outputs as part of $\mathbf{c}$. In our end-to-end system, we use two core sources during information collection:

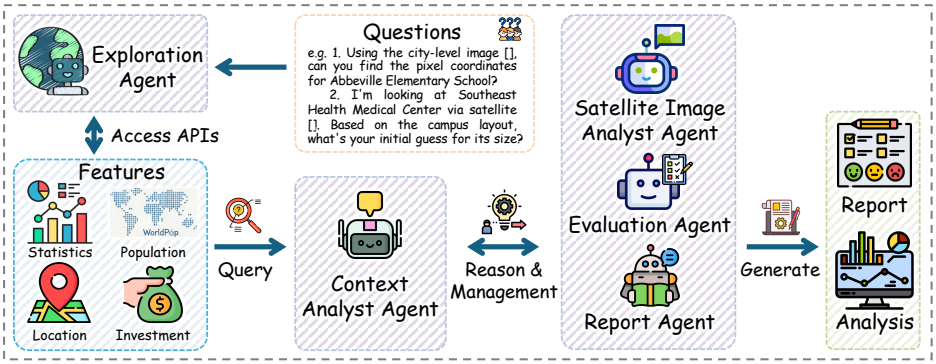


Figure 5: **Agentic geo-economic reasoning workflow**. Specialized stages select cached tools, analyze imagery and context, validate intermediate messages, forecast, and report.

(1) Google Maps Platform APIs Google (a;b), geocoding facility names, retrieving nearby POI inventories, and downloading map tiles for additional visual context. (2) WorldPop gridded population estimates Lloyd et al. (2017), obtain population density proxies around the target location, which provide a lightweight socioeconomic prior for demand forecasting. The data is precollected, and we simulate API access to achieve efficient execution, avoiding dependence on live API calls during evaluation while ensuring reproducible benchmark results.

Agentic Inference Pipeline. FORECAST synthesizes $\mathbf{v}$ and $\mathbf{u}$ and outputs a single structured message containing the final prediction:

$$\hat{y} = g_\theta(\mathbf{v}, \mathbf{u}), \qquad \hat{\mathbf{p}} = h_\theta(\mathbf{v}, \mathbf{u}), \tag{1}$$

where g_θ and h_θ are implemented via prompted VLM generation with strict formatting requirements. An ERROR handling mechanism will run at three checkpoints: it verifies that image paths are in-scope and requested cache entries that exist; when the requested features or the model output cannot be used in later round reasoning, the system will mark the message as error. It will also check that the final forecasts are parseable. This design prevents an API failure or hallucinated intermediate value from being treated as trusted in downstream.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We tested our agentic model with multiple open VLMs backbones including InternVL3 (8B, 14B), LLaVA-v1.6 (7B, 13B), Qwen2.5-VL (7B), and Qwen3.5-VL (9B). All models are evaluated under the same multi-round prompting protocol. For more evaluation details, please refer to Section 3.3.

4.2 DIRECT MODEL QUERY PERFORMANCE

We first evaluate the performance of directly querying models with the geo-eco question in our benchmark. Table 3 reports the quantitative performance across categories. **Err** refers to the error rate of numerical predictions, and **Fail** refers to the execution failure rate. It can be observed that numeric prediction remains challenging under multimodal visual and contextual information. Schools and libraries show high MAPE across models, and hospital financial indicators exhibit large MAE. In addition, output parse validity rate varies substantially for different models, indicating that models often fail to generate reliable output even before accuracy is considered. Across all categories, we observe two consistent patterns. First, even among parseable predictions, models frequently exhibit poor prediction results, producing values with large relative errors. Second, a lower parse rate in some settings suggests that stronger structured decoding and constraint enforcement are necessary for dependable evaluation and downstream use.

Table 3: Comparison between Agentic inference and Direct query across different base models and task domains. Values represent MAPE ($\times 100\%$) for Agentic and Direct modes, where $-$ indicates failed execution.

Model	Library		School		Hospital		House		Trail	
	Agentic	Direct	Agentic	Direct	Agentic	Direct	Agentic	Direct	Agentic	Direct
InternVL3-14B	3.04	**3.00**	**1.40**	1.64	**6.00**	10.00	1.97	**1.75**	**9.95**	-
InternVL3-8B	**1.99**	2.07	**1.96**	**1.96**	**9.38**	10.00	4.68	**3.15**	**9.69**	10.00
LLaVA-v1.6-13B	**3.41**	3.90	5.91	**5.68**	**5.47**	9.65	3.35	**1.75**	-	-
LLaVA-v1.6-7B	**3.46**	3.90	6.77	**4.94**	**6.21**	7.59	1.25	**0.00**	**6.60**	10.00
Qwen2.5-VL-7B	**2.38**	2.52	**1.66**	2.26	**5.97**	9.29	2.10	**1.75**	**8.81**	12.25
Qwen3.5-VL-9B	**2.10**	2.82	**1.25**	2.19	**1.26**	8.27	**1.36**	2.01	**6.51**	9.12

We visualize model capabilities across multiple dimensions with a radar chart in Fig. 6, which demonstrates the relative strengths and weaknesses of each model family in geo-economic reasoning, where larger values indicate better performance. Four axes summarize numeric prediction quality using MAPE for school, library, house, and trail, while the remaining axes capture localization precision as the within-threshold rate and output validity as the parse rate. The plot shows a consistent trade-off between task dimensions: models that are stronger on numeric prediction do not necessarily achieve the best localization, and several backbones cluster closely on parse rate despite diverging in forecasting performance. Overall, the figure highlights that current VLMs exhibit uneven strengths across geo-economic subtasks, with substantial headroom remaining in both calibrated numeric forecasting and city-scale visual grounding.

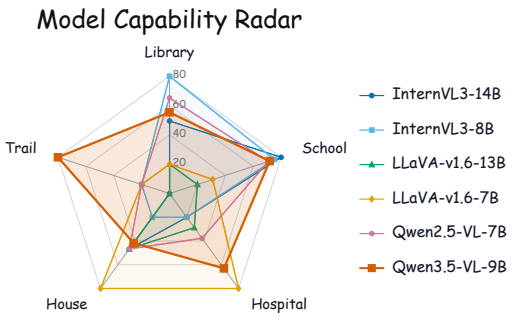


Figure 6: **Direct query model capability chart.** We visualize relative performance across six tasks with prediction accuracy.

4.3 Agentic Inference Results

Beyond direct prompting, our agentic pipeline decomposes each query into stages of exploration, visual analysis, context analysis, and forecasting, then emits structured numeric predictions and coordinates when required. We report the numerical prediction results against the direct query baseline on different base models. Table 3 compares multi-round *agentic mode* against a *direct query* baseline for numeric targets, reporting the error rate based on domain-specific tasks. Overall, agentic mode consistently outperforms direct query on numeric prediction tasks, which suggests that inferring with agents and leveraging intermediate tool outputs, such as POI counts or population, can improve performance on geo-economics-related prediction tasks. We also reported the results of location prediction in Appendix E.

We provide further analyses of inference efficiency, agentic localization behavior, failure cases, and dataset examples in Appendix D-G to better understand the practical characteristics and limitations of current models.

5 Conclusion

We have introduced GEBench, a benchmark for evaluating tool-augmented agents in geo-economic reasoning over heterogeneous geographic and socioeconomic information. The benchmark covers diverse real-world entities, including schools, hospitals, and libraries, and evaluates geo-economic analysis through an agentic and tool-interactive setting. GEBench provides a systematic testbed for studying multimodal grounding, evidence integration, and reasoning in geo-economic forecasting. A limitation of GEBench is that the benchmark targets are derived from public administrative records, which may contain reporting noise and temporal inconsistencies. We hope GEBench encourages future research on stronger retrieval, grounding, and reasoning modules, as well as more capable tool-using agents for real-world geo-economic analysis.

AI USE STATEMENT

Generative AI tools were used during research execution to assist with implementation and debugging of the experimental pipeline, including data processing, evaluation scripts, and experiment configuration. They were also used to generate synthetic benchmark instances following predefined task templates and source data. The generated samples were subsequently filtered and checked against the corresponding source information and task constraints before being included in the benchmark. Generative AI was not used to determine the main research hypothesis, define the core research direction, or design the key methodology of this work. All AI-assisted code, data, and experimental outputs were reviewed and verified by the authors. We take full responsibility for the final benchmark, experimental results, and conclusions reported in this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducible research, we provide detailed descriptions of benchmark construction, data processing procedures, evaluation protocols, and agent implementation details in the supplementary material. The GEBench dataset, evaluation scripts, and implementation code will be made publicly available after the review period.

To ensure consistent evaluation, all external tool responses, including geographic and socioeconomic information, are precollected and stored as fixed evidence rather than retrieved through live API calls during inference. This design avoids dependency on external service availability and enables reliable comparisons across different models under the same evaluation setting.

REFERENCES

Maheen Abbas and Fatima Ahmad. Spatialllm: Advancing urban spatial intelligence through multimodal large language models for classification, policy, and reasoning. *Frontiers in Computational Spatial Intelligence*, 2(2):54–62, 2024.

Agency for Healthcare Research and Quality. Hospital financial measures database (hfmd). `https://www.ahrq.gov/data/innovations/hfmd.html`.

Donghyun Ahn, Minhyuk Song, SeungEon Lee, Yubin Choi, Jihee Kim, Sangyoon Park, Hyunjoo Yang, and Meeyoung Cha. Fine-grained socioeconomic prediction from satellite images with distributional adjustment. In *CIKM*, pp. 3717–3721, 2023a.

Donghyun Ahn, Jeasurk Yang, Meeyoung Cha, Hyunjoo Yang, Jihee Kim, Sangyoon Park, Sungwon Han, Eunji Lee, Susang Lee, and Sungwon Park. A human-machine collaborative approach measures economic development using satellite imagery. *Nature Communications*, 14(1):6811, 2023b.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. *arXiv preprint arXiv:2505.20272*, 2026.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. Mindsearch: Mimicking human minds elicits deep AI searcher. In *ICLR*, 2025.

Zewen Chi, Li Dong, Qingxiu Dong, Yaru Hao, Xun Wu, Shaohan Huang, and Furu Wei. The era of agentic organization: Learning to organize with language models. *arXiv preprint arXiv:2510.26658*, 2025.

City of Austin . Trail counters daily totals in the austin. `https://catalog.data.gov/dataset/trail-counters-daily-totals`.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pp. 3213–3223, 2016.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 36:49250–49267, 2023.

Muhammad Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. In *ICCV*, pp. 7132–7142, 2025.

Jie Feng, Tianhui Liu, Yuwei Du, Siqi Guo, Yuming Lin, and Yong Li. Citygpt: Empowering urban spatial cognition of large language models. In *KDD*, pp. 591–602, 2025.

Google. Google maps platform: Places api (web service) documentation. `https://developers.google.com/maps/documentation/places/web-service`, a.

Google. Google maps platform: Maps static api documentation. `https://developers.google.com/maps/documentation/maps-static`, b.

Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *ICML*, volume 235, pp. 20413–20451, 2024.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. Mmsearch: Benchmarking the potential of large models as multi-modal search engines, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *ICLR*, 2024.

Varvara Krechetova and Denis Kochedykov. Geobenchx: Benchmarking llms for multistep geospatial tasks. *arXiv preprint arXiv:2503.18129*, 2025.

Mengke Li, Ming Li, Peng Zhang, Yan Wu, Wanying Song, and Lin An. Sar image change detection using pcanet guided by saliency detection. *GRSL*, 16(3):402–406, 2019.

Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. Econagent: large language model-empowered agents for simulating macroeconomic activities. In *Association for Computational Linguistics*, pp. 15523–15536, 2024.

Zekun Li, Jina Kim, Yao-Yi Chiang, and Muhao Chen. Spabert: A pretrained language model from geographic data for geo-entity representation. In *EMNLP*, volume EMNLP 2022 of *Findings of ACL*, pp. 2757–2769, 2022.

Zekun Li, Wenxuan Zhou, Yao-Yi Chiang, and Muhao Chen. GeoLM: Empowering language models for geospatially grounded language understanding. In *EMNLP*, pp. 5227–5240, 2023.

Zhonghang Li, Lianghao Xia, Xubin Ren, Jiabin Tang, Tianyi Chen, Yong Xu, and Chao Huang. Urban computing in the era of large language models. *ACM TIST*, 16(6):1–43, 2025.

Christopher T Lloyd, Alessandro Sorichetta, and Andrew J Tatem. High resolution global gridded data for use in population studies. *Scientific Data*, 4:170001, 2017.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025. ISSN 2731-5398.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024.

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

Mapbox. The mapbox direction api. `https://docs.mapbox.com`.

Fangyong Pan, Xinyi Huang, Yuxi Bi, Yunfan Gao, Yu Ye, and Haofen Wang. From tools to partners: How large language models are transforming urban planning. *AI Open*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Jonathan Roberts, Timo Lüddecke, Rehan Sheikh, Kai Han, and Samuel Albanie. Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. In *CVPRW*, pp. 554–563, 2024.

Ranjan Sapkota, Yang Cao, Konstantinos I Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025.

Simranjit Singh, Michael Fore, and Dimitrios Stamoulis. Geollm-engine: A realistic environment for building geospatial copilots. In *CVPRW*, pp. 585–594, 2024.

Yu-Zhu Sun, He-Li Sun, Jian-Cong Ma, Peng Zhang, and Xiao-Yong Huang. Multimodal agent ai: A survey of recent advances and future directions. *Journal of Computer Science and Technology*, 40 (4):1046–1063, 2025.

UK Government. Property transactions in the uk. `https://www.gov.uk/government/collections/property-transactions-in-the-uk`.

Ranga Raju Vatsavai. Geospatial foundation models: Recent advances and applications. In *BigSpatial*, pp. 30–33, 2024.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

Qihao Weng, Zhiwei Li, Yinxia Cao, Xiaoyan Lu, Paolo Gamba, Xiaoxiang Zhu, Yonghao Xu, Fan Zhang, Rongjun Qin, Micheal Y Yang, et al. How will ai transform urban observing, sensing, imaging, and mapping? *npj Urban Sustainability*, 4(1):50, 2024.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *NeurIPS*, 2024.

Liling Yang, Ning Chen, Jun Yue, Yidan Liu, Jiayi Ma, Pedram Ghamisi, Antonio Plaza, and Leyuan Fang. Survey of multimodal geospatial foundation models: Techniques, applications, and challenges. *arXiv preprint arXiv:2510.22964*, 2025.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583, 2020.

Zhenghang Yuan, Zhitong Xiong, Lichao Mou, and Xiao Xiang Zhu. Chatearthnet: a global-scale image–text dataset empowering vision–language geo-foundation models. *Earth System Science Data*, 2025.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, et al. The landscape of agentic reinforcement learning for llms: A survey. *arXiv preprint arXiv:2509.02547*, 2025a.

Qianheng Zhang, Song Gao, Chen Wei, Yibo Zhao, Ying Nie, Ziru Chen, Shijie Chen, Yu Su, and Huan Sun. Geoanalystbench: A geoai benchmark for assessing large language models for spatial analysis workflow and code generation. *TGIS*, 29(7):e70135, 2025b.

Yecheng Zhang, Huimin Zhao, and Ying Long. Cmab: A multi-attribute building dataset of china. *Scientific Data*, 12(1):430, 2025c.

Yifan Zhang, Jingxuan Li, Zhiyun Wang, Zhengting He, Qingfeng Guan, Jianfeng Lin, and Wenhao Yu. Geospatial large language model trained with a simulated environment for generating tool-use chains autonomously. *IJAEOG*, 136:104312, 2025d.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *TGRS*, 2024.

Mingming Zhao, Xiaokang Wei, Yuanqi Shao, Kaiwen Zhou, Lin Yang, Siwei Rao, Junhui Zhan, and Zhitang Chen. A^2Flow : automating agentic workflow generation via self-adaptive abstraction operators. *arXiv preprint arXiv:2511.20693*, 2025.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2026.

Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao, and Ranjay Krishna. Reinforced visual perception with tools. *arXiv preprint arXiv:2509.01656*, 2025.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Zheng Huang, Mingyu Liu, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, and Chunhua Shen. Active-o3: Empowering multimodal large language models with active perception via grpo, 2025.

Zillow Group. The zestimate api. `www.zillowgroup.com`.

Appendix for GEBench: Benchmarking Tool-Augmented Agents for Geo-Economic Analysis

A Dataset Composition and Geographic Coverage

This section reports the category composition and geographic coverage of GEBench. We distinguish generated dialogues from source records and unique coordinates because several question templates can be created from the same location. These statistics provide a clearer view of the dataset scale, spatial coverage, and potential sampling bias.

Table 4: Dataset composition. One source can be used by multiple question templates.

Category	Dialogues	Share	Sources	Unique coordinates	Primary geography
School	6,000	37.50%	5,380	5,288	USA and Australia
Hospital	3,000	18.75%	999	950	USA
Library	3,000	18.75%	18	18	Newcastle region, UK
House	3,000	18.75%	1,000	910	10 English cities
Trail	1,000	6.25%	39	30	Austin, USA
Total	16,000	100.00%	7,436	7,196	Three countries

Table 5: Country distribution. Dialogue counts include repeated use of source locations.

Country	Eligible sources	Generated dialogues
United States	5,451	9,033 (56.5%)
United Kingdom	1,018	6,000 (37.5%)
Australia	967	967 (6.0%)

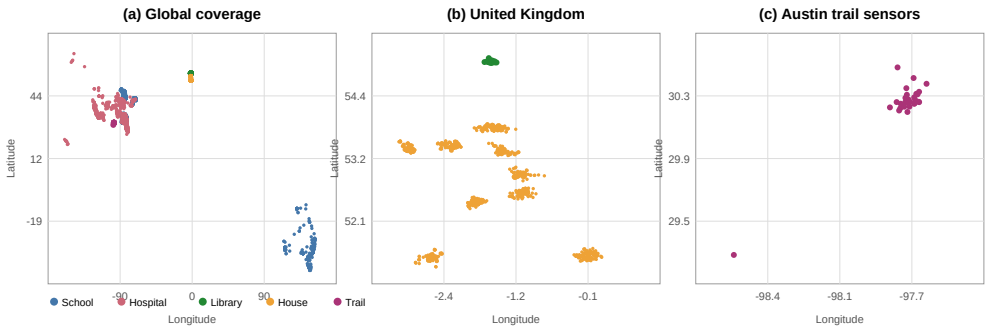


Figure 7: Spatial distribution of source locations. The data cover the United States, the United Kingdom, and Australia. The right panels enlarge the locations in the United Kingdom and the trail sensors in Austin.

Table 4 summarizes the composition of GEBench. Schools account for 37.5% of the dialogues. Hospitals, libraries, and houses each account for 18.75%, while trails account for 6.25%. Table 5 further summarizes the geographic coverage at the country level. Fig. 7 visualizes the finer-grained spatial distribution of source locations in GEBench. The locations are also unevenly distributed. The school data cover the United States and Australia, and the hospital data cover the United States. The library data are from the Newcastle region, the house data are from ten English cities, and the trail data are mainly from Austin.

Several question templates can be generated from one source location. Therefore, the number of dialogues is larger than the number of independent locations. This difference is largest for libraries and trails. Location is also related to category. For example, all hospital samples are from the United

States, while all house samples are from the United Kingdom. Performance across categories should therefore not be interpreted as country-level generalization. Future versions will consider geographic hold-out splits.

B Tool and Data Sources Summary

The external tools and data sources used to build our benchmark are summarized as follows:

1. **Google Maps Platform Places API** Google (a): used for geocoding and retrieving nearby-place inventories (*e.g.*, counts of restaurants, schools, hospitals) around a coordinate. In the benchmark, the derived POI-count features appear as structured fields and/or are embedded in the dialogue text.

2. **Google Maps Platform Static Maps API** Google (b): used to download place-level and city-level map tiles that serve as the primary visual information in GEM-QA.

3. **WorldPop population grids** Lloyd et al. (2017): used as an optional population-density prior during agentic information collection; the benchmark release does not require live access at evaluation time.

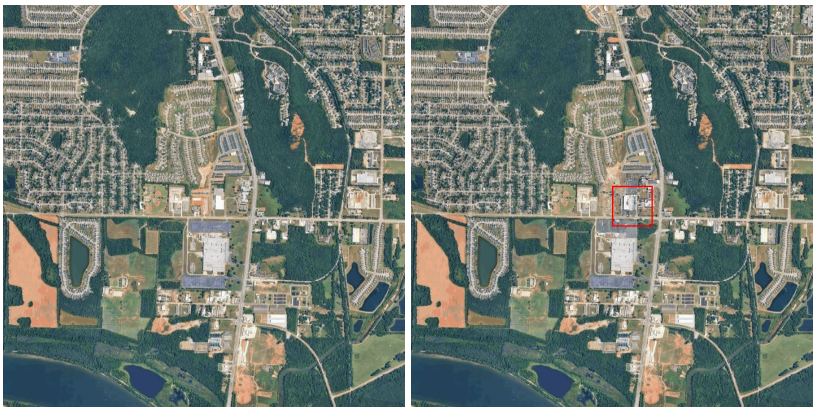
4. **AHRQ HFMD** Agency for Healthcare Research and Quality: provides hospital financial targets used in the hospital subset (e.g., revenue margin and net income ratio).

5. **UK Property Transaction** UK Government: provides the statistics on transactions for Stamp Duty Land Tax in England and Northern Ireland, Land and Buildings Transaction Tax in Scotland, and Land Transaction Tax in Wales. Specifically, we extract 1,000 transaction samples used in generating our QA benchmark.

6. **Austin Park Trail** City of Austin : This is a dataset that shows the twenty-four-hour counts from automated trail counter stations in the city of Austin. It gives a count of pedestrians and bicyclists who passed by the device location sensor.

C More Examples for Image Causal Removal

We present additional examples of edited satellite images for the image causal removal, as shown in Figs. 8 to 10. The results show that the edited image (Inpainting) maintains both global and local consistency, exhibiting a harmonious structure and a reasonable spatial distribution of the target buildings' removal. These edited images provide a basis for planning-oriented analysis.



Original *Inpainting*

Group 1: School

Figure 8: Examples of edited satellite images for schools.



Original *Inpainting*

Group 2: Hospital



Original *Inpainting*

Group 3: Library



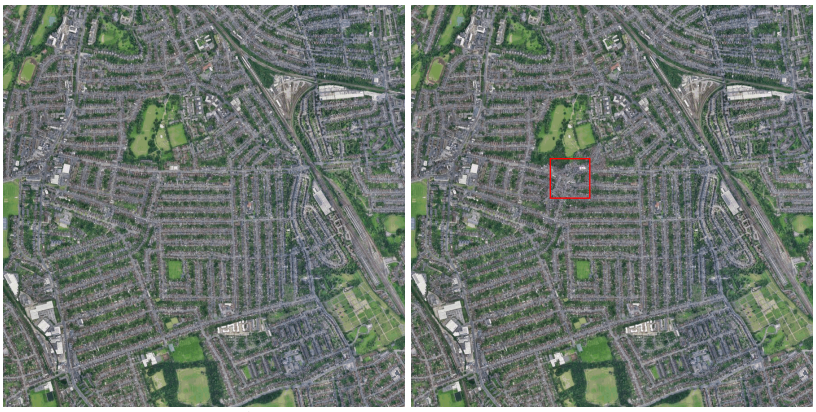
Original *Inpainting*

Group 4: Trail

Figure 9: Examples of edited satellite images for hospitals, libraries and trails.

D Inference Efficiency Analysis

Fig. 11 shows the end-to-end inference latency on the direct QA method, which evaluates the efficiency of single models. We observe a clear efficiency gap across backbones: InternVL3-8B is the fastest model at 1.50s per round, followed by InternVL3-14B at 2.22s per round. LLaVA-v1.6-7B and LLaVA-v1.6-13B run 5.20s and 7.97s, respectively, which are substantially slower. Overall, the



Original *Inpainting*

Group 5: House

Figure 10: Examples of edited satellite images for park trails and houses.

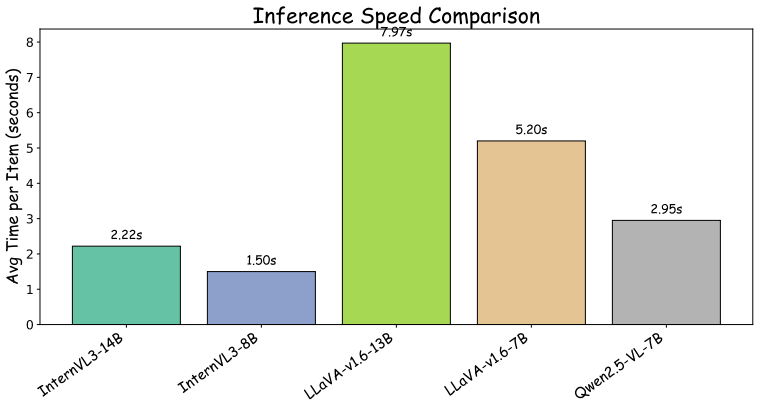


Figure 11: Inference latency comparison across evaluated backbones on the direct QA benchmark. We report average single-round inference time (lower is better).

slowest model is about 5.3× slower than the fastest, highlighting a practical trade-off between model capacity and inference time.

E AGENTIC LOCALIZATION PERFORMANCE

Table 6: Agentic localization accuracy reported by task categories. The agentic framework based on Qwen2.5-VL base model achieves the best performance.

Model	Library	School	Hospital	House	Trail
InternVL3-14B	17.23	7.61	8.33	4.50	0.00
InternVL3-8B	13.07	3.82	8.89	17.95	0.00
LLaVA-v1.6-13B	0.00	1.53	1.35	9.46	4.14
LLaVA-v1.6-7B	0.00	1.23	0.34	6.76	8.28
Qwen2.5-VL-7B	59.51	93.23	58.30	74.19	67.97

Table 6 reports localization performance as the percentage of predictions within 400 pixels. Here, agentic mode yields the most reliable improvements: decomposing perception into visual grounding and structured coordinate emission increases the fraction of valid, in-bounds outputs and reduces catastrophic format errors.

Despite these gains, localization remains far from solved. Within-threshold rates vary strongly by backbone and domain, indicating that city-scale grounding is still sensitive to prompt conventions and model-specific spatial priors. Overall, the numeric and localization results together highlight a key trade-off: agentic pipelines can stabilize how predictions are produced, but sustained improvements in what is predicted require stronger calibration and intermediate supervision.

F Failure case Analysis

We analyze failure cases from raw output answers. Two execution-time failures that dominate unusable predictions are focused on: **(i) extreme numeric outliers** and **(ii) output decoding failures** into the required schema.

In QA, we observe 185 decode failures where the parser cannot extract the required fields. Most of these failures are formatting issues rather than missing information. Models often return free-form text such as "Based on the satellite image…" instead of the required schema, produce malformed coordinate strings such as `SCHOOL_000917`, or output unresolved placeholders such as `<value> visitors` in `LIBRARY_002886` and `margin of and income ratio of` in `HOSPITAL_001755`. Another common case is a refusal-style response when the model claims the target is not visible in the image, as in `SCHOOL_005239`, which remains non-decodable under a strict coordinate parser.

In the agentic pipeline, decode failures mainly appear in localization as parsing failed in the output evaluation step. The rate varies sharply by backbone: InternVL3-8B accounts for most parsing failures, while InternVL3-14B has none in this run. These failures reflect strict output validation, where small deviations from the expected "(x, y)" format, or additional surrounding text, cause the evaluator to reject the prediction.

Even when decoding succeeds, a substantial portion of predictions are dominated by extreme outliers caused by scale mismatch. In QA numeric items, 112 cases exceed 500% MAPE, concentrated in school and library forecasting. A common pattern is a small ground truth paired with a typical large prediction. For `SCHOOL_005243`, the ground-truth student count is 5, yet multiple models predict 1,100–1,500, yielding MAPE up to 29,900%. Some library items collapse to round-number priors such as 1,000,000 visitors in `LIBRARY_002391`, producing multi-thousand-percent errors. These cases indicate that when the prompt provides weak or ambiguous scale cues, models default to coarse priors and fail to correct magnitude.

In the agentic benchmark, the failures are concentrated in trail and hospital domains. For hospitals, the most severe cases occur when the target is a small ratio near zero. For `HOSPITAL_001860`, the ground truth is 0.0003, but the prediction is 1,500,000.0, leading to a huge percentage error before clipping. Across models, many clipped failures repeat a small set of constants, such as over 100K, which suggests anchoring to default magnitudes and weak unit and scale control.

Across both settings, decode failures are mainly caused by strict parsers combined with unconstrained generation, where minor formatting deviations make otherwise reasonable responses unusable. Extreme outliers are mainly caused by poor unit and scale grounding, where models fall back to a few default magnitudes that can be far from the true scale. In practice, these findings motivate constrained decoding with validation and retry for formatting, and explicit unit and scale normalization or bounded prediction ranges for numeric forecasting.

G Representative GEBench Samples

To illustrate the diversity of our benchmark in real-world, satellite-driven reasoning tasks, we provide representative dataset examples in Fig. 12–18. These samples span both estimation and comparison problems across multiple urban entities and scales, including estimating the number of students and teachers in a school, comparing house values, estimating the value of an individual house, comparing the relative scale of two schools, estimating the location of a hospital, and estimating the number of visitors to a park trail and a library. These figures together can demonstrate the breadth of geographic contexts and the variety of quantitative and relational inferences required by our dataset.

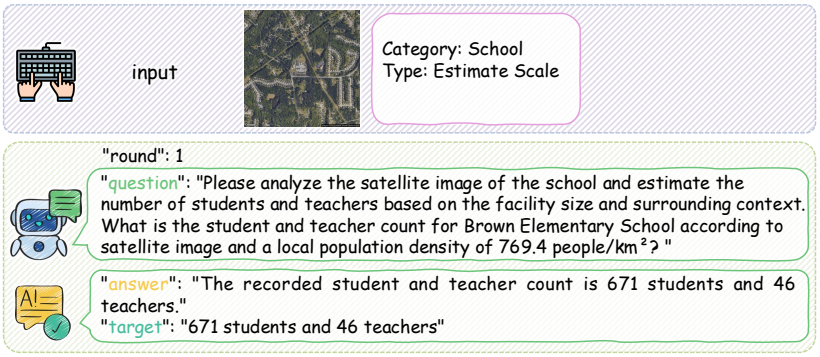


Figure 12: A dataset sample for estimating the number of students and teachers in a school.

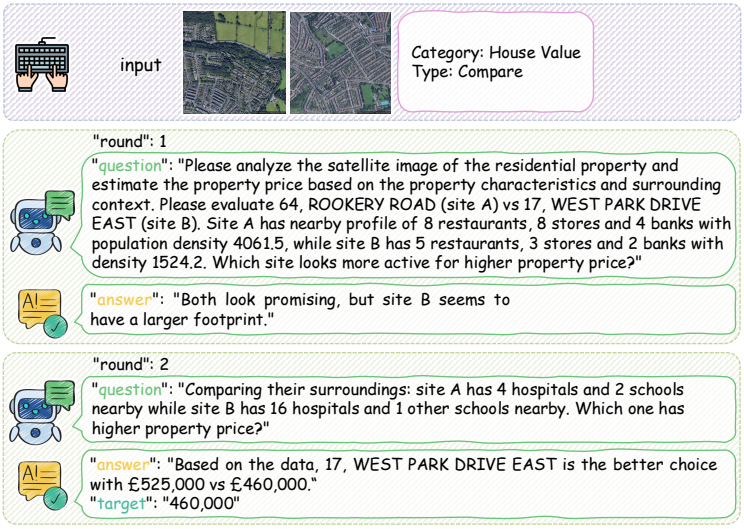


Figure 13: A dataset sample comparing house values.

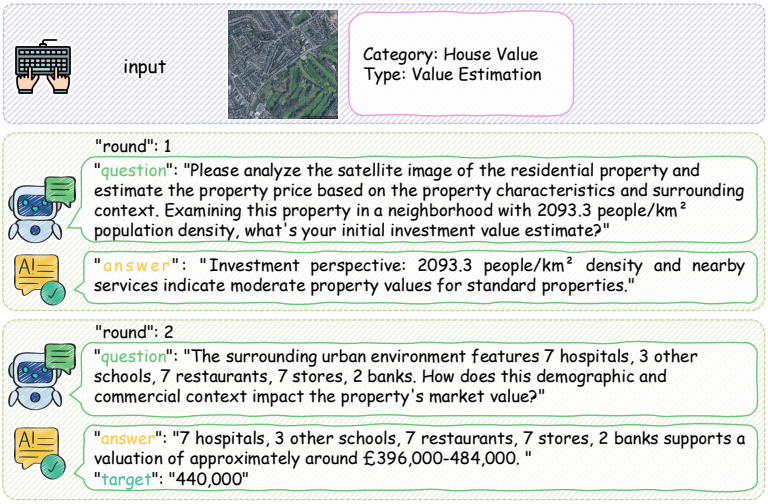


Figure 14: A dataset sample of estimating the value of a house.

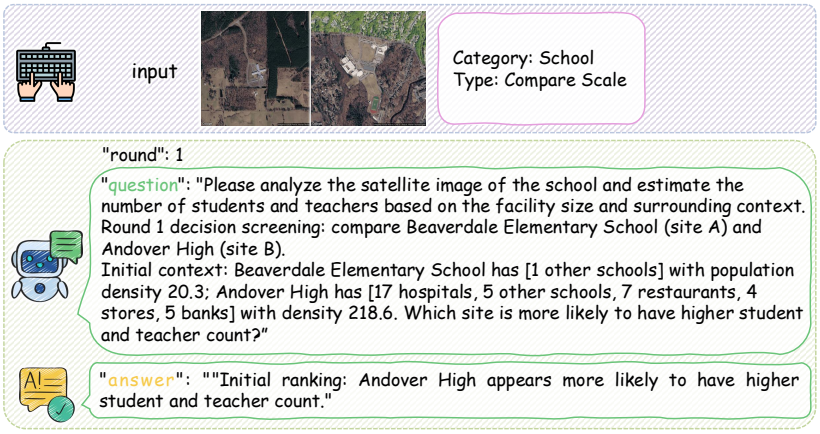


Figure 15: A dataset sample of comparing the scales of two schools.

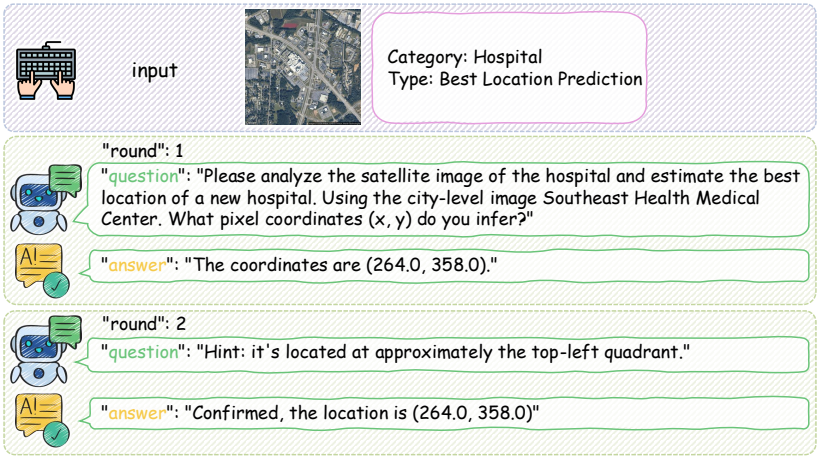


Figure 16: A dataset sample of estimating the location of a hospital.

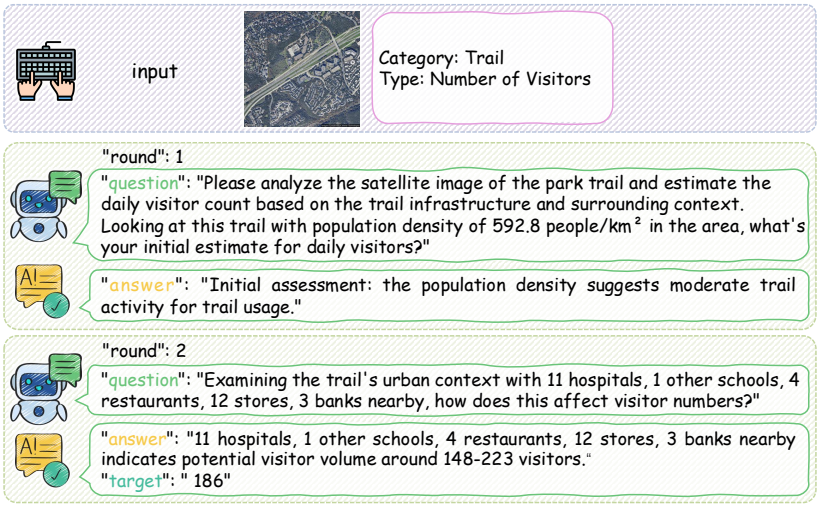


Figure 17: A dataset sample of estimating the number of visitors to a park trail.

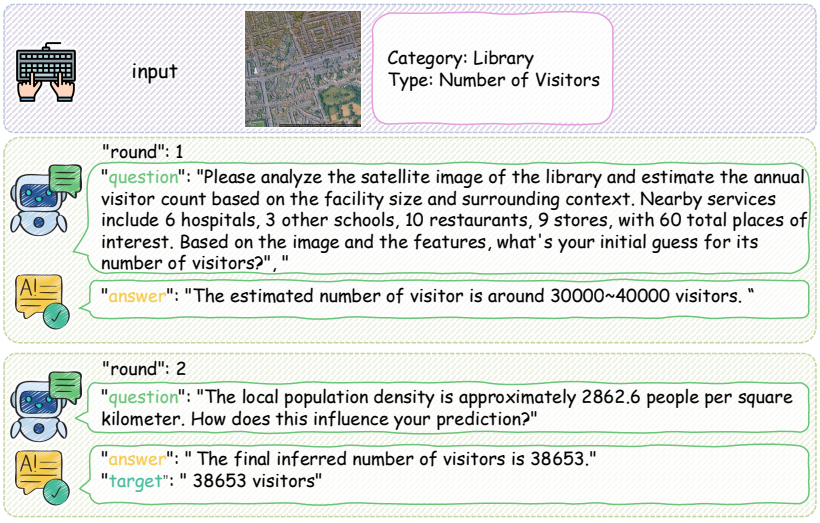


Figure 18: A dataset sample of estimating the number of visitors to a library.

H HOUSE TRANSACTION DATA DISTRIBUTION

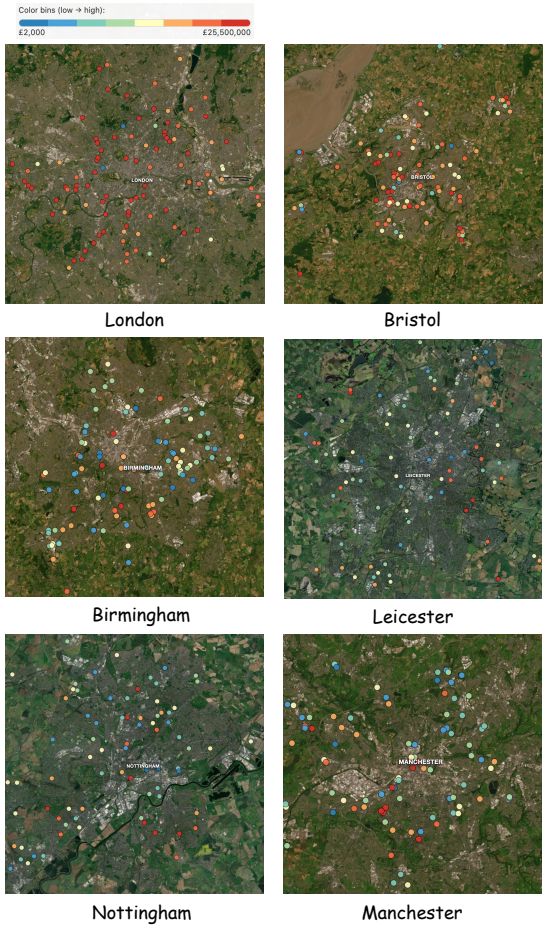


Figure 19: Visualization of the geographical distribution of house transaction prices used in our dataset.

Fig. 19 visualizes the spatial distribution of house transaction prices across the UK cities used to build our dataset. Each point represents a sampled transaction located by its geographic coordinates on satellite imagery and colored by its price level, which reveals both fine-grained variation within each city and systematic differences across cities. A clear insight is that London exhibits a markedly higher price regime than the other cities, with many more high-priced transactions and a broader spread, while cities such as Bristol, Birmingham, Leicester, Nottingham, and Manchester show comparatively lower and more concentrated price patterns. Overall, the visualization summarizes the dataset coverage and highlights the substantial price heterogeneity that motivates our value-oriented reasoning tasks.