DUDA-BENCH: BENCHMARKING LLM AGENTS ON MULTIMODAL DATA-DRIVEN URBAN DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Urban diagnosis integrates heterogeneous observations to identify urban problems, localize affected areas, and investigate contributing factors, informing evidence-based urban planning and management. However, its reliance on labor-intensive, case-specific expert workflows limits scalability and reuse, motivating the exploration of agent-based execution. To evaluate this capability, we introduce **DUDA-Bench**, a hierarchical and interactive benchmark that formalizes data-driven urban diagnosis as a multi-stage agent workflow. It comprises 86 atomic and 22 workflow tasks spanning four analytical stages, grounded in multimodal data from 12 cities covering five urban problem types. Evaluations of seven backbone models and five agent systems reveal a substantial gap between isolated analytical competence and end-to-end diagnosis, with system benefits varying across backbones. Trajectory analysis shows that unresolved evidence gaps propagate across stages, while successful recovery involves revising assumptions and actions using feedback. These findings highlight limitations in coordinating analytical capabilities across stages, particularly adaptive planning, evidence integration, and verification. More broadly, DUDA-Bench provides a framework for translating expert analytical workflows into hierarchical agent tasks and process-aware evaluation, supporting systematic assessment of end-to-end analytical capabilities.

1 INTRODUCTION

Urban diagnosis is the systematic assessment of urban conditions to identify problems and investigate their underlying causes, thereby informing urban planning decisions (Ye et al., 2021). Data-driven studies have examined urban sprawl, heat islands, and accessibility using heterogeneous spatial, temporal, and socioeconomic evidence (Shrestha et al., 2012; Xie et al., 2018; Wei & Sobrino, 2024; Kang & Lee, 2025). These analyses rely on experts to select evidence and iteratively refine interpretations, making them labor-intensive to reproduce across regions and problem types. LLM agents can reason over multiple steps and interact with external tools (Yao et al., 2023), offering new opportunities for data-driven urban diagnosis. This motivates evaluating their ability to carry out complete diagnostic workflows and identifying the capabilities that limit success.

Assessing these capabilities requires evaluation settings that capture the analytical demands of urban diagnosis. Existing urban and geospatial benchmarks assess interactive decision making, multimodal understanding, and executable spatial analysis (Feng et al., 2025b;a; Yu et al., 2026), while general and scientific agent benchmarks evaluate multi-step tool use and trajectory-level execution (Wang et al., 2026; Shen et al., 2026; Liu et al., 2026a). These efforts provide relevant foundations, but leave open how effectively agents can sustain an evidence-driven diagnostic process across heterogeneous urban observations(Table 1). Unlike a predefined operation such as indicator estimation, open-ended diagnosis requires agents to determine what evidence is needed, acquire and integrate observations, and refine conclusions through interaction, as illustrated by the urban-heat example in Fig. 3 (Appendix A). Evaluating this process requires examining both diagnostic outcomes and intermediate analytical progress. The central challenge is to understand how individual analytical capabilities translate into coherent end-to-end diagnosis, and where their coordination breaks down.

To address this evaluation gap, we introduce **DUDA-Bench**, a benchmark for **data-driven urban diagnosis** that combines heterogeneous urban data, adaptive analytical workflows, and outcome-

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.

Table 1: Positioning comparison of DUDA-Bench with representative urban, geospatial, and Earth-observation benchmarks.

Category	Benchmark	Data Coverage						Agentic Workflow Tasks			Evaluation	
		Map Data	Urban Indicators	Remote Sensing	Street View	Temporal Data	Cities	Interactive Execution	Multi-stage Workflow	Planning	Outcome	Process-aware
Urban-agent benchmarks	CityBench Feng et al. (2025b)	✓	✓	✓	✓	✓	13	✓	✓	✗	✓	✗
	CityEQA Zhao et al. (2025)	✗	✗	✗	✓	✗	–	✓	✓	✓	✓	✗
	USTBench Lai et al. (2026)	✓	✓	✗	✗	✓	–	✓	✓	✗	✓	✓
Urban-multimodal benchmarks	UrBench Zhou et al. (2025)	✗	✗	✓	✓	✗	11	✗	✗	✗	✓	✗
	DynamicVL Xuan et al. (2026)	✗	✓	✓	✗	✓	42	✗	✗	✗	✓	✗
	UBench Feng et al. (2025a)	✓	✗	✓	✓	✗	3	✓	✓	✗	✓	✗
	CityLens Liu et al. (2026b)	✗	✓	✓	✓	✗	17	✗	✗	✗	✓	✗
Geospatial benchmarks	GEOBench-VLM Danish et al. (2025)	✗	✗	✓	✗	✓	–	✗	✗	✗	✓	✗
	Earth-Agent Feng et al. (2026)	✗	✓	✓	✗	✓	–	✓	✓	✓	✓	✓
	GeoAgentBench Yu et al. (2026)	✓	✗	✗	✗	✗	–	✓	✓	✓	✓	✓
Ours	**DUDA-Bench**	✓	✓	✓	✓	✓	**12**	✓	✓	✓	✓	✓

Note. Multi-temporal Data denotes repeated real-world observations over time. Planning denotes adaptive task decomposition and execution-path revision; Process-aware denotes evaluation of intermediate analytical progress. ✓: supported; ✗: not a primary component; –: not directly comparable.

and process-level evaluation. It comprises two complementary task types: **workflow tasks** assess complete diagnostic processes spanning problem framing, urban profiling, diagnostic reasoning, and validation and synthesis, while **atomic tasks** assess individual analytical operations extracted from these workflows (Fig. 1). Workflow tasks extend predefined urban operations to diagnosing what the problem is, where it occurs, and what evidence supports its explanation through iterative analysis (Fig. 3). Together, the two task types distinguish competence on individual operations from the ability to coordinate them into a complete diagnosis.

DUDA-Bench integrates three components: a **hierarchical task taxonomy** (§3.2) linking atomic operations to workflow tasks, a **multimodal urban case dataset** (§3.3) providing standardized data across cities and problem types, and an **agentic environment** (§3.4) supporting data inspection, domain tools, reusable skills, and code execution under multiple agent systems. Hierarchical outcome evaluation measures diagnostic success, while stage-aligned checkpoints assess intermediate analytical progress. Together, these components enable controlled comparisons and help locate where evidence acquisition, reasoning, and execution succeed or fail.

We evaluate seven backbone models under a shared execution framework and compare five agent systems on two backbones, complemented by trajectory analysis and component ablations. The results reveal a substantial capability gap between isolated analytical operations and end-to-end diagnosis, with agent-system benefits varying across backbones. Tracing workflow execution shows how this gap develops: unresolved evidence gaps persist into downstream reasoning, whereas successful recovery involves revising assumptions and actions in response to feedback. Together, these findings highlight limitations in coordinating analytical capabilities across stages, particularly adaptive planning, evidence integration, and verification. Component ablations further show that these capabilities depend on access to computational and domain-specific support during execution.

Accordingly, our main contributions are summarized as follows:

- We formulate **data-driven urban diagnosis** as an open-ended agentic task and establish a hierarchical task design linking individual analytical operations to complete diagnostic workflows. This formulation enables evaluation of the gap between component capabilities and end-to-end task completion.
- We build **DUDA-Bench** with standardized multi-modal urban cases, an interactive environment supporting domain tools, reusable skills, and code execution, and a protocol com-

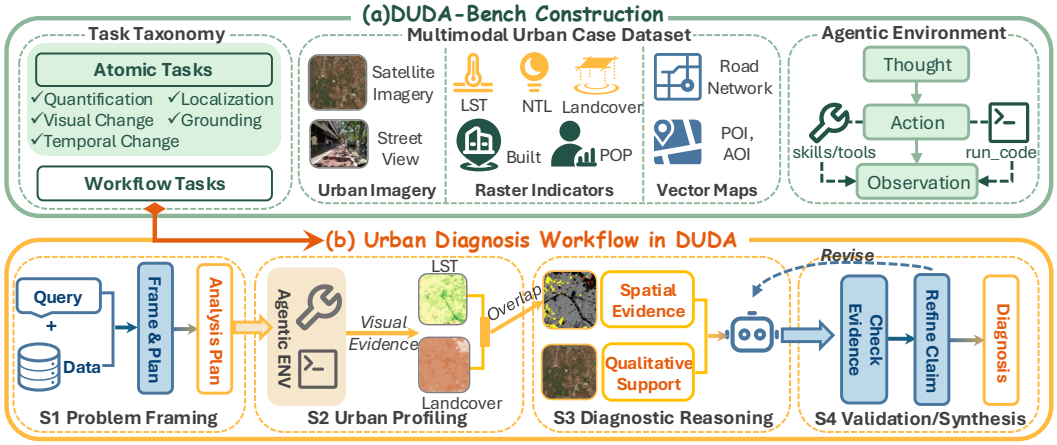


Figure 1: Overview of DUDA-Bench. The benchmark connects a hierarchical task taxonomy and multimodal urban case dataset with an interactive agent environment, and formulates urban diagnosis as a four-stage analytical workflow.

bining outcome assessment with stage-aligned checkpoints. These resources support evaluation across regions, problem types, and agent systems.

- We systematically evaluate backbone models and agent systems, revealing a substantial gap between atomic competence and workflow-level diagnosis. Trajectory analysis and component ablations highlight limitations in cross-stage capability coordination and the role of interactive execution in supporting diagnostic performance.

2 Related Work

Data-Driven Urban Diagnosis. Urban problem diagnosis provides an analytical basis for understanding urban conditions and informing interventions (Gu et al., 2018; Bibri, 2021; Foster, 2016). Existing data-driven studies typically address individual problems in specific regions, including urban sprawl (Shrestha et al., 2012), shrinkage (Xie et al., 2018), heat islands (Wei & Sobrino, 2024), and accessibility (Kang & Lee, 2025). These studies establish the conceptual and empirical foundations of urban diagnosis, but do not formulate it as a unified task across problem types, regions, and heterogeneous data sources. Building on this foundation, DUDA-Bench formalizes data-driven urban diagnosis as a common, evaluable agent task.

Urban and Agent Benchmarks. Existing urban benchmarks evaluate complementary capabilities, including interactive urban agents (Feng et al., 2025b; Zhao et al., 2025; Lai et al., 2026), multimodal urban understanding (Zhou et al., 2025; Feng et al., 2025a; Xuan et al., 2026; Liu et al., 2026b), and executable geospatial or Earth-observation analysis (Danish et al., 2025; Feng et al., 2026; Yu et al., 2026). In parallel, general and scientific agent benchmarks increasingly evaluate multi-step execution and trajectories rather than only final answers (Yao et al., 2025; Wang et al., 2026; He et al., 2026; Shen et al., 2026; Liu et al., 2026a; Varambally et al., 2026). DUDA-Bench builds on these complementary directions by firstly formulating and benchmarking **data-driven urban diagnosis** as a unified agentic task, requiring agents to integrate heterogeneous urban evidence, execute multi-stage analyses interactively, and be evaluated at both outcome and process levels.

3 DUDA-Bench

DUDA-Bench is a hierarchical and interactive benchmark for data-driven urban diagnosis, built around three components: a task taxonomy, a multimodal urban dataset, and an agentic environment. Construction proceeds through four stages. **Literature-Grounded Task Construction** translates

expert diagnostic workflows into workflow tasks and constituent atomic operations. **Multimodal Data Collection and Organization** grounds these tasks in observations from 12 cities covering five urban problem types over a ten-year period. **Interactive Environment Construction** makes these data accessible through tools, skills, and code execution, enabling agents to carry out the tasks through traceable multi-turn interaction. Finally, **Hierarchical Evaluation** combines task-outcome assessment with stage-aligned checkpoints to evaluate both individual analytical capabilities and the execution of complete diagnostic workflows.

3.1 Task Formulation

In DUDA-Bench, we formulate data-driven urban diagnosis as a multi-step, tool-assisted task $\mathcal{T} = (q, \mathcal{D}, \mathcal{E})$, where q specifies the diagnostic objective and spatiotemporal scope, $\mathcal{D}$ denotes the available multimodal urban data, and $\mathcal{E}$ is an executable environment providing data access and analytical tools. Given $\mathcal{T}$, an agent iteratively gathers and analyzes evidence, with intermediate results informing subsequent actions and revisions. The goal is to produce an evidence-grounded diagnosis that identifies the urban problem, localizes affected areas, and explains contributing factors.

3.2 Task Taxonomy

Task construction. We construct tasks in two steps. First, two urban-domain experts, each with over five years of research experience, manually select and review urban diagnosis studies (Gu et al., 2018; Bibri, 2021; Foster, 2016; Shrestha et al., 2012; Xie et al., 2018; Wei & Sobrino, 2024; Kang & Lee, 2025). These studies cover five typical scenarios grouped into three categories: urban morphology (urban sprawl and shrinkage), eco-climatic resilience (urban heat islands and green-space fragmentation), and service accessibility (accessibility inequality). The experts translate the studies' diagnostic objectives into workflow tasks and extract their reported conclusions as reference ground truth. Second, we extract individual analytical operations from these workflows and formulate atomic tasks according to five capability types: indicator quantification, temporal change analysis, spatial problem localization, visual change interpretation, and evidence-to-metric grounding. Atomic reference outputs are computed using the benchmark dataset and manually checked. Tasks at both levels are reviewed for data availability and evaluation feasibility.

Atomic-to-Workflow Task Taxonomy. The resulting taxonomy comprises **22 workflow tasks** and **86 atomic tasks**, with their composition and distribution summarized in Table 4 in Appendix A.1. Workflow tasks follow four analytical stages (Fig. 1): **(S1) Problem Framing** defines the target issue and spatiotemporal scope; **(S2) Urban Profiling** gathers multimodal observations and computes indicators; **(S3) Diagnostic Reasoning** integrates evidence to identify problems and interpret possible mechanisms; and **(S4) Validation and Synthesis** checks evidence, revises claims, and synthesizes the diagnosis. Across these scenarios, workflow tasks are grouped into three complexity levels according to the number of urban issues within the same geographic unit: single-issue, dual-issue, and multi-issue tasks involve one, two, and three or more issues, respectively. Atomic tasks assess individual operations within this structure, while workflow tasks require agents to select and coordinate these capabilities through open-ended, multi-turn analysis. This paired design supports assessment of both individual analytical competence and end-to-end diagnostic performance. Further construction details are provided in Appendix A.1.

3.3 Multimodal Urban Case Dataset

Dataset scope. We construct a multimodal urban dataset to support cross-region and cross-problem evaluation of data-driven urban diagnosis. Rather than combining existing benchmark datasets, we collect observations directly from original data providers and standardize them under consistent spatial scopes, temporal organization, and annotation criteria. The dataset covers two spatial levels: city-level observations capture broad urban change, while study-region observations focus on case-specific areas identified from the literature and support finer-grained localization. It contains data for **12 cities** and **22 study regions** over **2015–2024**, spanning North America, Europe, and Asia. Totally, our dataset contains 7,071 data assets and 1,507,693 data records. Dataset statistics and composition are summarized in Table 5 in Appendix A.2.

Data collection and standardization. The city pool and data requirements are grounded in empirical studies of the five target urban problem types (Shrestha et al., 2012; Xie et al., 2018; Wei & Sobrino, 2024; Kang & Lee, 2025). We assemble complementary urban data from eight sources: OSM, WorldPop, GHSL, VIIRS, Landsat 8/9, ESA WorldCover, Sentinel-2 L2A, and Mapillary. Each source is clipped or queried for the predefined city or study region and organized with harmonized spatial and temporal metadata. Table 5 summarizes the dataset's coverage, composition, and scale, while Table 6 details source contents and access channels. Both of them appear in Appendix A.2.

3.4 Agentic Environment

Interactive execution. DUDA provides a unified interactive environment for urban-diagnosis agents implemented with different systems, including Direct prompting, Plan-and-Execute (Wang et al., 2023), ReAct (Yao et al., 2023), MAF-Magentic (Fourney et al., 2024), and LAMBDA (Sun et al., 2025). The environment supports multi-turn interaction through standardized interfaces to urban data, domain-specific tools, reusable skills, and a code sandbox. Agents can iteratively inspect data, compute indicators, validate intermediate results, and revise their analysis based on accumulated evidence. All systems share the same task inputs and tool interfaces, supporting controlled comparisons of backbone models and agent systems while allowing each system to follow its own execution strategy.

Domain tools, skills, and data access. The environment provides 7 domain skills and 8 tool families to support urban diagnosis. Skills offer reusable guidance on data selection, indicator definition, and spatiotemporal comparison, while tools enable data retrieval, computation, visualization, and code execution. Through these interfaces, agents access evidence as *numerical values*, *visual evidence*, and *textual descriptions*. Visual evidence includes both source imagery and visualizations generated from raster data, allowing agents to combine quantitative analysis with visual interpretation. For analyses beyond the predefined tools, agents can use `run_code` to process data and implement custom computations in the sandbox. The complete mapping between domain skills, tool families, and executable operations is provided in Appendix A.4.

3.5 Evaluation Protocol

Outcome Evaluation. We evaluate atomic operations and complete diagnostic workflows separately. Atomic outputs are checked by task-specific validators for numerical accuracy, spatial agreement, directional correctness, or structural validity. A task passes only if its output meets the specified criterion. Workflow evaluation requires more than a correct issue label: agents must also locate the problem and provide quantitative support. Following urban-analysis practices that characterize problems through their definition, spatial pattern, and diagnostic indicators (Zhang et al., 2023; Liu et al., 2023; Deilami et al., 2018), we assess three aspects: **Recognition** (A), identification of the target issue family; **Localization** (B), coverage of the reference locations; and **Indicator F1** (C), precision and recall of the reported diagnostic evidence.

For a single run on a workflow target, let f be the target issue family and $\widehat{\mathcal{F}}$ the predicted families. Localization and evidence are evaluated using only predictions associated with f. We compute:

$$A = \mathbf{1}\{f \in \widehat{\mathcal{F}}\}, \tag{1}$$

$$B = A \cdot \mathrm{Coverage}(\widehat{\mathcal{L}}, \mathcal{L}), \tag{2}$$

$$C = A \cdot \frac{2|\widehat{\mathcal{E}} \cap \mathcal{E}|}{|\widehat{\mathcal{E}}| + |\mathcal{E}|}, \tag{3}$$

$$S = A \cdot \frac{B + C}{2}. \tag{4}$$

Here, $\mathcal{L}$ and $\widehat{\mathcal{L}}$ denote the reference and predicted locations associated with the target issue. $\mathrm{Coverage} \in [0, 1]$ measures the proportion of reference locations covered by the prediction, using spatial overlap or matched areas of interest. Where an area constraint applies, predictions covering more than 25% of the city's bounding box receive zero localization credit, preventing agents from gaining credit through overly broad spatial predictions. $\mathcal{E}$ and $\widehat{\mathcal{E}}$ denote the required and reported

diagnostic evidence sets. Incorrect issue recognition yields zero localization, indicator, and overall scores.

We first average scores across repeated runs of each task or target, then macro-average across atomic tasks or workflow targets:

$$\text{Atomic PR} = \frac{1}{N_a}\sum_{i=1}^{N_a}\bar{p}_i, \qquad \text{Workflow Score} = \frac{1}{N_w}\sum_{i=1}^{N_w}\bar{S}_i, \tag{5}$$

where PR denotes pass rate, $\bar{p}_i$ is the mean binary pass outcome for atomic task i, $\bar{S}_i$ is the mean continuous score for workflow target i, and N_a and N_w are their respective counts.

Checkpoint-Driven Trajectory Evaluation. Outcome metrics indicate whether an agent completes the diagnosis but not where the execution diverges. Inspired by GTA-2 (Wang et al., 2026), we therefore decompose each workflow into 14 verifiable checkpoints aligned with the four analytical stages. Each checkpoint specifies an analytical sub-goal rather than a prescribed tool sequence, and GLM-5.3 (GLM-5-Team et al., 2026) judges the observable trajectory as *pass*, *partial*, or *fail*. Representative checkpoint results are reported in Fig. 2(B4). We further validate the judge on 56 randomly sampled ReAct trajectories (716 checkpoint verdicts): two independent annotators achieve 93.6% agreement, while GLM-5.3 reaches 90.4% accuracy against adjudicated human labels. Additional verification details are provided in Appendix A.3.

4 Experiments

We conduct experiments to address three research questions(RQs) that progressively examine overall performance, execution dynamics, and capability bottlenecks in multimodal urban diagnosis. **RQ1:** How do current LLM/MLLM agents perform on multimodal data-driven urban diagnosis, and how does performance vary across backbone models and agent execution strategies? **RQ2:** How do failures and recoveries unfold over multi-stage diagnostic workflows, and what trajectory-level behaviors distinguish successful from unsuccessful executions? **RQ3:** Which analytical capabilities are most critical to successful end-to-end urban diagnosis, and how do weaknesses in these capabilities manifest in workflow-level performance? Our experiments follow this progression, beginning with overall performance and controlled backbone and agent-system comparisons, then examining variation across benchmark factors, and finally tracing workflow executions to identify failure patterns and the capabilities underlying them.

4.1 Experimental Setup

Models and Agent Systems. We organize our experiments into three complementary settings to disentangle the effects of backbone capability, agent-system design, and individual execution components. First, for the *backbone comparison*, we evaluate seven representative frontier and open-weight models: GPT-5.6-Terra (OpenAI, 2026), GPT-5.6-Sol (OpenAI, 2026), Gemini-3.7-Flash (Google DeepMind, 2026), Claude-Sonnet-5 (Anthropic, 2026), Kimi-K3 (Team et al., 2026), Qwen3.8-Max and Qwen3.8-27B (Qwen Team, 2026). To isolate backbone capability from agent scaffolding, all models are evaluated under the same ReAct-style execution framework (Yao et al., 2023), which supports multi-turn reasoning and interaction with a code sandbox, domain-specific tools, and reusable skills. Each backbone is evaluated on the full benchmark comprising 22 workflow tasks and 86 atomic tasks. Second, for the *agent-system comparison*, we fix the backbone to Qwen3.8-Max and Qwen3.8-27B and compare five execution paradigms: Direct prompting, Plan-and-Execute, ReAct, MAF-Magentic based on the Magentic multi-agent orchestration paradigm, and LAMBDA, a specialized multi-agent system for data analysis. This comparison examines how different levels of planning, iterative environment interaction, multi-agent coordination, and domain-specific agent design affect urban diagnostic performance. Finally, we conduct a *component ablation* using Qwen3.8-Max with ReAct as the base configuration, comparing the full system against three variants that remove code execution (*w/o code*), skill routing (*w/o skill-routing*), or domain tools (*w/o tools*), respectively, to quantify the contribution of each execution component.

Metrics. We report **Atomic PR** for atomic-task completion and four workflow-level outcome metrics: **Recognition**, **Localization**, **Indicator F1**, and the aggregated **Workflow Score**, as defined

Table 2: Main backbone results on DUDA-Bench under the ReAct system.

Model	Atomic PR	Workflow Outcome				Efficiency	
		Recognition	Localization	Indicator F1	Workflow Score	Avg. Turns	Tokens (K)
GPT-5.6-Terra	0.419	0.250	0.080	0.063	0.071	**9.9**	184.7
Gemini-3.7-Flash	**0.558**	0.432	0.169	0.195	0.182	15.9	239.1
Claude-Sonnet-5	**0.558**	0.386	0.159	0.111	0.135	18.2	228.5
GPT-5.6-Sol	0.500	0.318	0.080	0.103	0.091	12.1	**177.9**
Kimi-K3	0.465	0.591	0.080	0.189	0.134	13.4	196.8
Qwen3.8-Max	0.500	**0.727**	**0.274**	**0.281**	**0.277**	16.7	244.5
Qwen3.8-27B	0.500	0.432	0.125	0.141	0.133	20.3	322.3

Notes. Atomic PR and Workflow PR denote the corresponding task-level performance metrics. Avg. Turns and Tokens (K) report the mean number of agent turns and total token usage (in thousands) per workflow episode. Repetitions are averaged within each problem, and problems are equally weighted. Best and second-best distinct results are shown in **bold** and underlined, respectively; lower is better for efficiency metrics.

Table 3: Comparison of agent systems on DUDA-Bench using Qwen3.8-Max and Qwen3.8-27B backbones.

Agent System	Recognition	Localization	Indicator F1	Workflow Score
Qwen3.8-Max				
Direct	0.659	0.216	0.126	0.171
Plan-and-Execute	0.591	0.125	0.128	0.126
ReAct	**0.727**	**0.274**	**0.281**	**0.277**
MAF-Magentic	0.295	0.057	0.081	0.069
LAMBDA	0.227	0.125	0.070	0.097
Qwen3.8-27B				
Direct	**0.500**	0.148	0.101	0.124
Plan-and-Execute	0.364	0.091	0.087	0.089
ReAct	0.432	0.125	0.141	0.133
MAF-Magentic	0.432	0.091	0.096	0.093
LAMBDA	**0.500**	**0.181**	**0.149**	**0.165**

Notes. Repetitions are averaged within each problem, and problems are equally weighted. Best and second-best distinct results within each backbone are shown in **bold** and underlined, respectively.

in Section 3.5. To characterize execution efficiency, we additionally report the average number of agent turns (**Avg. Turns**) and total token usage in thousands (**Tokens (K)**) per workflow episode. Higher values are better for all outcome metrics, whereas lower Avg. Turns and Tokens indicate greater efficiency.

4.2 Overall Performance and Agent System Effects (RQ1)

End-to-end diagnosis remains substantially harder than isolated analytical tasks. Table 2 and Fig. 2(B2) show that Qwen3.8-Max achieves the highest Workflow Score (0.277), while the remaining models range from 0.071 to 0.182. By contrast, Atomic PR varies within a narrower range of 0.419–0.558. Notably, Qwen3.8-Max, GPT-5.6-Sol, and Qwen3.8-27B all obtain an Atomic PR of 0.500 but diverge markedly at the workflow level. This suggests that competence on individual operations is insufficient for completing a multi-stage diagnostic process. Table 3 further shows that this gap depends on the execution system: Qwen3.8-Max performs best with ReAct, whereas Qwen3.8-27B performs best with LAMBDA.

Tool-use strategies expose different evidence bottlenecks. Execution traces reveal three broad styles: code-dominant behavior for Gemini-3.7-Flash and Qwen3.8-27B, high-level tool reliance for Claude-Sonnet-5 and GPT-5.6-Sol, and a more mixed strategy for Qwen3.8-Max. Code enables finer-grained spatial comparisons and specialized indicators, but introduces more execution failures, while high-level tools are more stable but tend to remain at aggregate statistics. Visual tools are rarely used across models, consistent with the weak visual-change capability in Fig. 2(B6). Im-

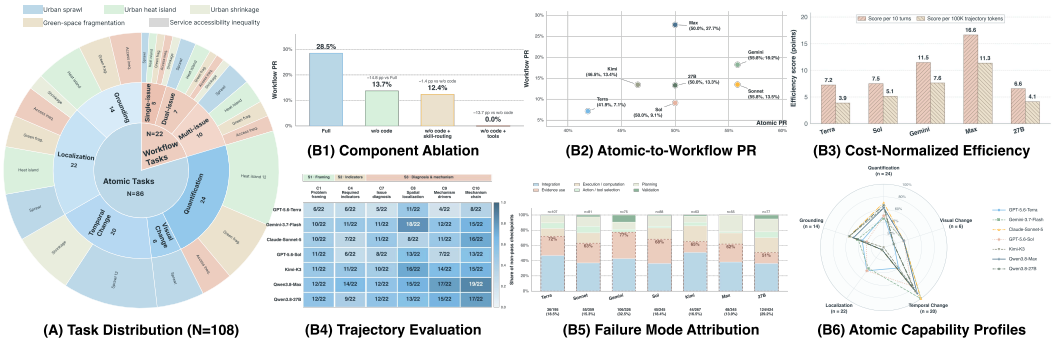


Figure 2: Benchmark composition and experimental analyses. (A) Task distribution across workflow and atomic tasks. (B1–B6) Component ablation, atomic-to-workflow performance, cost-normalized efficiency, trajectory evaluation, failure-mode attribution, and atomic capability profiles.

portantly, code frequency alone does not predict success: Gemini-3.7-Flash and Qwen3.8-27B are both code-heavy, yet their Workflow Scores differ substantially. What matters is whether tool use produces evidence that directly supports the diagnostic objective.

System effects are backbone dependent, and higher cost does not imply better performance. As shown in Table 3, the relative advantage of Qwen3.8-Max over Qwen3.8-27B changes across systems and even reverses under LAMBDA, indicating a clear backbone–system interaction. Efficiency exhibits the same non-monotonic pattern. Table 2 reports 9.9–20.3 turns and 177.9K–322.3K tokens per workflow, while Fig. 2(B3) shows that Qwen3.8-Max achieves the strongest cost-normalized performance. Qwen3.8-27B consumes the largest interaction budget without attaining the highest Workflow Score. Effective diagnosis therefore depends more on how computation is allocated to evidence acquisition and analysis than on interaction budget alone.

4.3 Failure Propagation and Recovery Patterns (RQ2)

Early evidence gaps tend to persist into downstream diagnosis. Figure 2(B4) shows that checkpoint failures are not confined to the final reasoning stage: agents often formulate plausible analysis plans but fail to convert required indicators into usable observations. This gap then propagates from profiling to diagnosis, producing trajectories in which a missing spatial or quantitative analysis is later reported as insufficient evidence rather than actively resolved. For example, GPT-5.6-Sol can identify the relevant modalities while still omitting discriminative indicators, whereas Qwen3.8-Max more consistently carries intermediate evidence into mechanism-level reasoning. The central difficulty is therefore not merely planning what to inspect, but maintaining alignment between the target problem, the evidence collected, and the downstream diagnostic claim.

Successful execution does not guarantee valid evidence. Fig. 2(B5) attributes checkpoint failures to execution, tool selection, evidence use, planning, and validation. Runtime errors make failures explicit and can guide recovery, but analytically invalid outputs may go unnoticed. For example, Qwen3.8-27B sometimes uses inappropriate land-cover categories or produces grid statistics dominated by missing values despite successful execution. Without checks on category meaning, coordinate systems, and data validity, these outputs can mislead subsequent analysis. Repeated calls can still be productive when failures reveal new information about the data. Verification must therefore assess whether outputs support the analysis, beyond whether tools execute successfully.

Successful recovery involves revising the analysis in response to errors. Two trajectories illustrate this pattern. In a successful Qwen3.8-27B heat-analysis run, errors prompt the agent to check how raster coordinates map to geographic locations, which pixels contain valid data, and which years are available. These checks guide subsequent analysis and allow the agent to complete a localized comparison. In an unsuccessful sprawl run with a similar number of turns, the agent repeatedly encounters alignment and coding errors but continues without revising its assumptions, ultimately failing to produce reliable spatial evidence. Table 3 likewise shows that additional interaction oppor-

tunities do not consistently improve performance across backbone–system combinations. Together, these observations suggest that effective recovery requires agents to use error feedback to revise assumptions and adjust subsequent actions, while additional turns alone do not ensure success.

4.4 Capability Bottlenecks and Interactive Affordances (RQ3)

The main bottleneck is maintaining executable evidence goals across stages. Figure 2(B4) shows a clear gap between identifying relevant evidence and carrying it through to mechanism-level reasoning: agents often know which modalities to inspect, yet fail to convert missing indicators into concrete follow-up actions. This reflects three coupled capability requirements: (1) maintaining which diagnostic subgoals remain unresolved, (2) preserving valid data and execution state across tool calls, and (3) integrating intermediate evidence into a consistent downstream explanation. The atomic profiles in Fig. 2(B6) reinforce this view: relatively strong quantification or temporal analysis does not compensate for weaknesses in localization, visual interpretation, or evidence grounding. End-to-end diagnosis therefore depends on composing these capabilities rather than excelling at any single operation.

Model scale does not directly determine realizable workflow capability. Qwen3.8-27B illustrates how external feedback can compensate for weaker base capability. Although it matches several stronger models on Atomic PR in Table 2, its workflow performance varies substantially across systems in Table 3, and it even surpasses Qwen3.8-Max under LAMBDA. Successful trajectories show that executable feedback—such as array shapes, missing-value patterns, and tool errors—helps the model revise local assumptions and preserve partial evidence. Conversely, stronger models can still fail when they do not convert recognized uncertainty into further analysis. GPT-5.6-Sol, for example, often identifies missing evidence and states limitations correctly, yet does not always turn these gaps into targeted follow-up computations. This contrast suggests that interactive execution primarily expands *realizable* capability by improving evidence acquisition and control, rather than simply reflecting model size.

Supplementary component ablations expose where execution capability is lost. Figure 2(B1) shows that removing code reduces Workflow Score from 28.5% to 13.7%; disabling skill routing in the no-code setting further lowers it to 12.4%, while removing domain tools reduces performance to zero. These components support different parts of the workflow: code enables finer-grained spatial and multimodal analysis, skill routing helps select and invoke relevant procedural guidance, and domain tools provide direct access to standardized evidence. The result indicates that analytical intent alone is insufficient unless the agent can translate it into executable evidence.

5 Conclusion

We introduce **DUDA-Bench**, a benchmark that formulates data-driven urban diagnosis as a unified agentic task spanning problem framing, urban profiling, diagnostic reasoning, and validation. DUDA-Bench combines hierarchical atomic and workflow tasks, a multimodal urban dataset, an interactive agentic environment, and outcome- and process-level evaluation to assess complete analytical workflows. Experiments across multiple backbone models and agent systems reveal a substantial gap between isolated analytical competence and end-to-end diagnosis, with recurring failures in evidence acquisition, state maintenance, recovery and cross-stage integration. At the same time, trajectory and component analyses show that verifiable interaction, code execution, skill routing, and domain tools can help agents better realize their analytical capabilities. We hope DUDA-Bench provides a useful foundation for evaluating and improving multimodal agents toward more reliable, evidence-grounded urban analysis.

AI use statement

In this work, we used generative AI tools to translate and refine author-prepared text, assist literature discovery, and discuss and refine research ideas and methodology grounded in our review of prior work. The authors determined the substantive content and organization of the manuscript and manually checked retrieved publications and references against their original sources. Codex and Claude assisted with code implementation and debugging. LLMs were also used as evaluated agents and for AI-assisted judging, as described in the evaluation protocol(Section 3.5). We did not use generative

AI to synthesize the underlying urban observations, which originate from real-world data sources documented in Appendix A.2; mathematical proof assistance is not applicable to this work. The authors reviewed all AI-assisted work and take responsibility for the final text, code, analyses, claims, and other artifacts.

Ethics statement

DUDA-Bench uses real-world urban data from publicly accessible sources documented in Appendix A.2. Our dataset release will comply with the applicable source licenses and terms of use, preserve required attribution, and redistribute only materials for which redistribution is permitted; otherwise, we will provide source references and instructions for authorized access. Released materials will be reviewed for privacy risks, with potentially identifying content excluded or appropriately protected. Geographic coverage and source-data biases may affect benchmark results and limit their generalization. DUDA-Bench is intended for research evaluation, and agent-generated diagnoses should not be used for consequential planning decisions without expert review and local validation.

Reproducibility statement

The benchmark construction section describes the task hierarchy, multimodal data organization, interactive environment, and evaluation protocol, while the experimental setup specifies the evaluated models, agent systems, and execution settings. Appendix A.2 documents the data sources and dataset composition. Upon acceptance, we will publicly release the benchmark code and dataset, subject to the source-specific redistribution conditions described in the Ethics statement. The release will include data collection and preprocessing scripts, task definitions, evaluation code, experiment configurations, and a README with setup and execution instructions, supporting reproduction of the experiments and extension of the data collection pipeline to additional cities.

References

Anthropic. Introducing claude sonnet 5. https://www.anthropic.com/news/claude-sonnet-5, June 2026. Accessed: 2026-09-23.

Simon Elias Bibri. Data-driven smart sustainable cities of the future: An evidence synthesis approach to a comprehensive state-of-the-art literature review. *Sustainable Futures*, 3:100047, 2021. ISSN 2666-1888. doi: https://doi.org/10.1016/j.sftr.2021.100047. URL https://www.sciencedirect.com/science/article/pii/S266618882100006X.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7132–7142, October 2025.

Kaveh Deilami, Md. Kamruzzaman, and Yan Liu. Urban heat island effect: A systematic review of spatio-temporal factors, data, methods, and mitigation measures. *International Journal of Applied Earth Observation and Geoinformation*, 67:30–42, 2018. ISSN 1569-8432. doi: https://doi.org/10.1016/j.jag.2017.12.009. URL https://www.sciencedirect.com/science/article/pii/S0303243417302994.

Jie Feng, Shengyuan Wang, Tianhui Liu, Yanxin Xi, and Yong Li. Urbanllava: A multi-modal large language model for urban intelligence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6209–6219, October 2025a.

Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. Citybench: Evaluating the capabilities of large language models for urban tasks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 5413–5424, New York, NY, USA, 2025b. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737375. URL https://doi.org/10.1145/3711896.3737375.

Peilin Feng, Zhutao Lv, Junyan Ye, Xiaolei Wang, Xinjie Huo, Jinhua Yu, Wanghan Xu, Wenlong Zhang, LEI BAI, Conghui He, and Weijia Li. Earth-agent: Unlocking the full landscape of earth observation with agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dkIXAbWuxO.

Kelleann Foster. Geodesign parsed: Placing it within the rubric of recognized design theories. *Landscape and Urban Planning*, 156:92–100, 2016. ISSN 0169-2046. doi: https://doi.org/10.1016/j.landurbplan.2016.06.017. URL https://www.sciencedirect.com/science/article/pii/S0169204616301244. Geodesign—Changing the world, changing design.

Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Erkang, Zhu, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, Peter Chang, Ricky Loynd, Robert West, Victor Dibia, Ahmed Awadallah, Ece Kamar, Rafah Hosn, and Saleema Amershi. Magentic-one: A generalist multi-agent system for solving complex tasks, 2024. URL https://arxiv.org/abs/2411.04468.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Google DeepMind. Gemini 3.7 flash model card. https://deepmind.google/models/model-cards/gemini-3-7-flash/, August 2026. Accessed: 2026-09-23.

Yexuan Gu, Brian Deal, and Linda Larsen. Geodesign processes and ecological systems thinking in a coupled human-environment context: An integrated framework for landscape architecture. *Sustainability*, 10(9), 2018. ISSN 2071-1050. doi: 10.3390/su10093306. URL https://www.mdpi.com/2071-1050/10/9/3306.

Pengfei He, Zhenwei Dai, Bing He, Hui Liu, Xianfeng Tang, Hanqing Lu, Juanhui Li, Jiayuan Ding, Subhabrata Mukherjee, Suhang Wang, Yue Xing, Jiliang Tang, and Benoit Dumoulin. TRAJECT-bench:a trajectory-aware benchmark for evaluating agentic tool use. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TZWnWvsQ0X.

Sunju Kang and Gunwon Lee. Assessing accessibility and equity in childcare facilities through 2sfca: Insights from housing types in seongbuk-gu, seoul. *ISPRS International Journal of Geo-Information*, 14(7), 2025. ISSN 2220-9964. doi: 10.3390/ijgi14070247. URL https://www.mdpi.com/2220-9964/14/7/247.

Siqi Lai, Yansong Ning, Zirui Yuan, Zhixi Chen, and Hao Liu. USTBench: Benchmarking and dissecting spatiotemporal reasoning capabilities of LLMs as urban agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ETzBStUFJy.

Fan Liu, Xiaozhao Zeng, and Hao Liu. Towards multimodal data-driven scientific discovery powered by LLM agents. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=kZHSvETWdi.

Tianhui Liu, Hetian Pang, Xin Zhang, Tianjian Ouyang, Zhiyuan Zhang, Jie Feng, Yong Li, and Pan Hui. Citylens: Evaluating large vision-language models for urban socioeconomic sensing. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=kswX9NfAlo.

Yang Liu, Mei-Po Kwan, Man Sing Wong, and Changda Yu. Current methods for evaluating people's exposure to green space: A scoping review. *Social Science & Medicine*, 338:116303, 2023. ISSN 0277-9536. doi: https://doi.org/10.1016/j.socscimed.2023.116303. URL https://www.sciencedirect.com/science/article/pii/S0277953623006603.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026. Accessed: 2026-09-23.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Yujiong Shen, yajie yang, Zhiheng Xi, Binze Hu, Huayu Sha, Qiyuan Peng, Jiazheng Zhang, Junlin Shang, Jixuan Huang, Yutao Fan, Jingqi Tong, Ming Zhang, Shihan Dou, Zhenfei Yin, Xingjun Ma, LEI BAI, Tao Gui, Qi Zhang, Xuanjing Huang, and Yu-Gang Jiang. Sciagentgym: Benchmarking multi-step scientific tool-use in LLM agents. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=0Moj0YgFEF.

Milan K. Shrestha, Abigail M. York, Christopher G. Boone, and Sainan Zhang. Land fragmentation due to rapid urbanization in the phoenix metropolitan area: Analyzing the spatiotemporal patterns and drivers. *Applied Geography*, 32(2):522–531, 2012. ISSN 0143-6228. doi: https://doi.org/10.1016/j.apgeog.2011.04.004. URL https://www.sciencedirect.com/science/article/pii/S0143622811000658.

Maojun Sun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang. Lambda: A large model based data agent. *Journal of the American Statistical Association*, 121 (553):1–13, July 2025. ISSN 1537-274X. doi: 10.1080/01621459.2025.2510000. URL http://dx.doi.org/10.1080/01621459.2025.2510000.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, Peng Chen, Ruijue Chen, Wentao Chen, Xin Chen, Yang Chen, Yanru Chen, Yifei Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Dazhi Cheng, Yean Cheng, Jialei Cui, Jingbing Cui, Anqi Dai, Jiaqi Deng, Hao Ding, Rui Ding, Shaofeng Ding, Mengfan Dong, Mengnan Dong, Yuhao Dong, Yuxin Dong, Angang Du, Chenzhuang Du, Dikang Du, Jusen Du, Yulun Du, Yu Fan, Jing Feng, Qiulin Feng, Yichen Feng, Kelin Fu, Qiang Fu, Fuxuan Gao, Hongcheng Gao, Jingyue Gao, Tong Gao, Weijia Gao, Shangyi Geng, Jie Gong, Linhu Gong, Shengao Gong, Xiaochen Gong, Qizheng Gu, Yicheng Gu, Shuhao Guan, Haiqing Guo, Shiqi Guo, Xiang Guo, Zhengyan Guo, Beixi Hao, Wenxin Hao, Xiaoru Hao, Dailan He, Haotian He, Lehan He, Qi He, Weiran He, Xinran He, Xinyi He, Yibo He, Yunjia He, Chao Hong, Tiange Hong, Hao Hu, Jiaxi Hu, Ruikun Hu, Weiming Hu, Yangyang Hu, Zhenxing Hu, Liang Hua, Jinbin Huang, Ke Huang, Ruiyuan Huang, Siying Huang, Weixiao Huang, Yan Huang, Zhengjie Huang, Zhiqi Huang, Yulong Hui, Chaobo Jia, Yutong Jiang, Zhejun Jiang, Zuoyou Jiang, Wenyi Jin, Xinyi Jin, Yu Jing, Huanjun Kong, Guokun Lai, Aidi Li, Cheng Li, Chengyuan Li, Cong Li, Fang Li, Guanyu Li, Haoyang Li, Jia Li, Junxiong Li, Lei Li, Letian Li, Lincan Li, Weihong Li, Wentao Li, Xintong Li, Yang Li, Yishen Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Zhengxiao Li, Zhiyuan Li, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zichao Lin, Ziyan Lin, Bill Liu, Boxiao Liu, Chuan Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu,

Yanming Liu, Yibo Liu, Yipeng Liu, Zhengying Liu, Zhiheng Liu, Enzhe Lu, Haoyu Lu, Linqiang Lu, Tingzhan Lu, Zhiyuan Lu, Aotian Luo, G. Luo, Junyu Luo, Yifan Luo, B. Lyu, Wenzhou Lyu, Shaoguang Mao, Yuan Mei, Xin Men, Minqing Ni, Yixuan Niu, Siyuan Pan, Shujun Peng, Zhangyang Qi, Ruoyu Qin, ZeChao Qin, Zeyu Qin, Haiquan Qiu, Jianxin Qiu, Jiezhong Qiu, Bowen Qu, Yuhao Qu, Zeyu Shang, Youbo Shao, Han Shen, Jincheng Shi, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Wingchun Siu, Pengwei Song, Xiaoxi Song, Jianlin Su, Yunfeng Su, Zhaochen Su, Lin Sui, Jingsong Sun, Junyao Sun, Shaoning Sun, Shuzhe Sun, Tongyu Sun, Yujun Sun, Yunpeng Tai, Chuning Tang, Heyi Tang, Sirui Tang, Zecheng Tang, Chaoran Tian, Rongpeng Tian, Yu Tian, Wei Tu, Chensi Wang, Chuang Wang, Chunjie Wang, Dinglu Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hao Wang, Huaqing Wang, Hui Wang, Jiayi Wang, Jinglong Wang, Jinhong Wang, Jiuzheng Wang, Linian Wang, Shaobo Wang, Shenzhi Wang, Shuyi Wang, Si Wang, Siyuan Wang, Tianfu Wang, Wenjue Wang, Xingran Wang, Xinmei Wang, Xinyuan Wang, Xusheng Wang, Yalin Wang, Yangkun Wang, Yao Wang, Yaoyu Wang, Yejie Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhenhao Wang, Zhongsheng Wang, Zifan Wang, Chu Wei, Ming Wei, Shouxin Wei, Zichen Wen, Fan Wu, Haoning Wu, Rucong Wu, Wenhao Wu, Xiaoxue Wu, Yingcong Wu, Yongqi Wu, Yuxin Wu, Zijian Wu, Xinglang Xian, Chenxuan Xiang, Yuye Xiang, Bocheng Xiao, Chenjun Xiao, Xin Xiao, Jin Xie, Xiaotong Xie, Yifeng Xie, Zhe Xie, Bowei Xing, Yiming Xiong, Baosheng Xu, Boyu Xu, Jiale Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Qingtao Xu, Shuyao Xu, Suting Xu, Tiantian Xu, Tianxiang Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ye Xu, Yueni Xu, Ziyao Xu, Haonan Xue, Junjie Yan, Yaoyao Yan, Fan Yang, Guangyao Yang, Hao Yang, Junwei Yang, Ruoyu Yang, Wenjie Yang, Xiaofei Yang, Xinyu Yang, Yi Yang, Yiling Yang, Ying Yang, Yuchen Yang, Zhen Yang, Zhilin Yang, Zian Yang, Zuhao Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhanbo Ye, Bohong Yin, Haoxiang Yin, Xietong Yin, Chengzhen Yu, Haozhen Yu, Longhui Yu, Shengnan Yu, Shuying Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Tongtian Yue, Wei Yue, Yang Yue, Dunyuan Zha, Haobing Zhan, B. H. Zhang, Dehao Zhang, Fei Zhang, Hao Zhang, Haoyuan Zhang, Huanyu Zhang, Jiapei Zhang, Jiaxuan Zhang, Jin Zhang, Kaiyi Zhang, Miaozhen Zhang, Puqi Zhang, Qinglei Zhang, Rong Zhang, Rui Zhang, Shaoshuai Zhang, Shiyi Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yangkun Zhang, Ye Zhang, Yichi Zhang, Yikun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Zijing Zhang, Bin Zhao, Chenguang Zhao, Feifan Zhao, Jinglun Zhao, Jinxiang Zhao, Shuai Zhao, Wenshuo Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Haozhi Zheng, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Haofeng Zhong, Lei Zhong, Longguang Zhong, M. Zhou, Qiankang Zhou, Runjie Zhou, Ruozhang Zhou, Xinyu Zhou, Yiqiao Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yangjunfeng Zhu, Yuxuan Zhu, Zhen Zhu, Chen Zhuang, Weiyu Zhuang, and Xinxing Zu. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Sumanth Varambally, Marshall Fisher, Jas Thakker, Yiwei Chen, Zhirui Xia, Yasaman Jafari, Ruijia Niu, Manas Jain, Veeramakali Vignesh Manivannan, Zachary Novack, Luyu Han, Srikar Eranky, Salva Rühling Cachay, Taylor Berg-Kirkpatrick, Duncan Watson-Parris, Yian Ma, and Rose Yu. Zephyrus: An agentic framework for weather science. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=aVeaNahsID.

Jize Wang, Xuanxuan Liu, Yining Li, Songyang Zhang, Yijun Wang, Zifei Shan, Xinyi Le, Cailian Chen, Xinping Guan, and Dacheng Tao. Gta-2: Benchmarking general tool agents from atomic tool-use to open-ended workflows, 2026. URL https://arxiv.org/abs/2604.15715.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models, 2023. URL https://arxiv.org/abs/2305.04091.

Letian Wei and José A. Sobrino. Surface urban heat island analysis based on local climate zones using ecostress and landsat data: A case study of valencia city (spain). *International Journal of Applied Earth Observation and Geoinformation*, 130:103875, 2024. ISSN 1569-8432. doi: https://doi.org/10.1016/j.jag.2024.103875. URL https://www.sciencedirect.com/science/article/pii/S1569843224002292.

Yichun Xie, Hongmian Gong, Hai Lan, and Shi Zeng. Examining shrinking city of detroit in the context of socio-spatial inequalities. *Landscape and Urban Planning*, 177:350–361,

2018. ISSN 0169-2046. doi: https://doi.org/10.1016/j.landurbplan.2018.03.002. URL https://www.sciencedirect.com/science/article/pii/S0169204618300719.

Weihao Xuan, Junjue Wang, Heli Qi, Zihang Chen, Zhuo Zheng, Yanfei Zhong, Junshi Xia, and Naoto Yokoya. DynamicVL: Benchmarking multimodal large language models for dynamic city understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=zubCrOvUZ4.

Zherui Yang, Fan Liu, Yansong Ning, and Hao Liu. Evods: Self-evolving autonomous data science agent with skill learning and context management. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 6128–6139, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3818002. URL https://doi.org/10.1145/3770855.3818002.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. {τ}-bench: A benchmark for \underline{T}ool-\underline{A}gent-\underline{U}ser interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Zhongnan Ye, Hanxue Wei, and Chuanren Lin. Community renewal with urban diagnosis: Bajiao community, shijingshan district, beijing. In *Proceedings of the 57th ISOCARP World Planning Congress*, 2021. URL https://isocarp.org/app/uploads/2022/02/ISOCARP_2021_Ye_74.pdf.

Bo Yu, Cheng Yang, Dongyang Hou, Chengfu Liu, Jiayao Liu, Chi Wang, Zhiming Zhang, Haifeng Li, and Wentao Yang. Geoagentbench: A dynamic execution benchmark for tool-augmented agents in spatial analysis, 2026. URL https://arxiv.org/abs/2604.13888.

Pan Zhang, Debarchana Ghosh, and Sohyun Park. Spatial measures and methods in sustainable urban morphology: A systematic review. *Landscape and Urban Planning*, 237:104776, 2023. ISSN 0169-2046. doi: https://doi.org/10.1016/j.landurbplan.2023.104776. URL https://www.sciencedirect.com/science/article/pii/S0169204623000956.

Yong Zhao, Kai Xu, Zhengqiu Zhu, Yue Hu, Zhiheng Zheng, Yingfeng Chen, Yatai Ji, Chen Gao, Yong Li, and Jincai Huang. CityEQA: A hierarchical LLM agent on embodied question answering benchmark in city space. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12465–12480, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.630. URL https://aclanthology.org/2025.emnlp-main.630/.

Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. Urbench: a comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i10.33163. URL https://doi.org/10.1609/aaai.v39i10.33163.

A Benchmark Construction Details

A.1 Task Construction

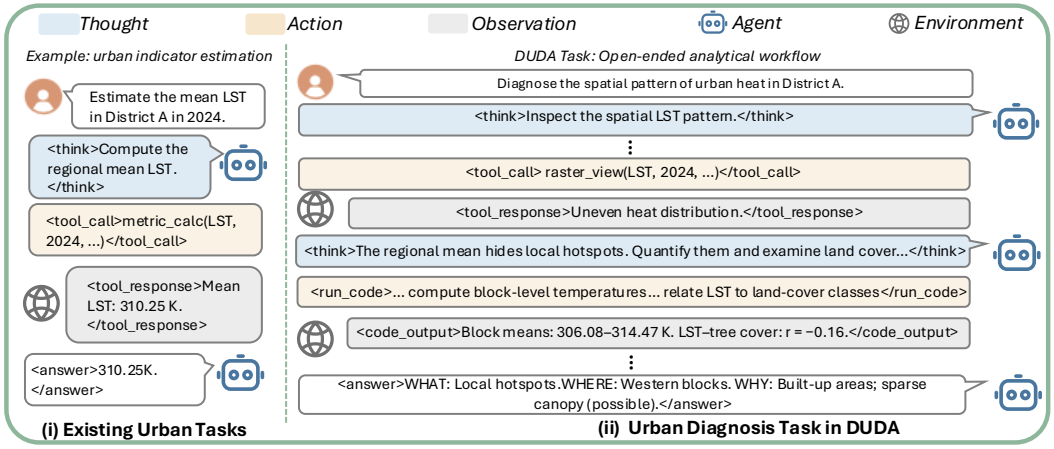


Figure 3: From predefined urban analysis to open-ended urban diagnosis. Existing urban tasks typically evaluate a specified analytical operation, whereas DUDA requires an agent to iteratively determine, acquire, and integrate evidence throughout a diagnostic workflow.

Task illustration. Fig. 3 contrasts a predefined urban indicator estimation task with a DUDA urban diagnosis task. While the existing tasks ask the agent to perform a specified calculation, urban diagnosis requires it to inspect spatial patterns, select analytical steps, integrate tool outputs, and produce a localized, evidence-grounded diagnosis.

Task taxonomy. Table 4 summarizes the composition of DUDA-Bench: 22 workflow tasks categorized by problem complexity and urban problem type, and 86 atomic tasks categorized by capability type and urban problem type.

Table 4: Task composition and distribution of DUDA-Bench.

Workflow Tasks ($N = 22$)		**Atomic Tasks** ($N = 86$)	
Problem Complexity (3)	***N***	***Capability Types (5)***	***N***
Single-issue	5	Indicator quantification	32
Dual-issue	7	Temporal change analysis	20
Multi-issue	10	Spatial problem localization	22
		Visual change interpretation	6
		Evidence-to-metric grounding	6

Problem Types (5)	***Workflow N***	***Atomic N***
Urban Morphology: Urban sprawl	6	24
Eco-Climatic Resilience: Urban heat island	6	24
Eco-Climatic Resilience: Green-space fragmentation	4	16
Service accessibility inequality	4	12
Urban Morphology: Urban shrinkage	2	10

A.2 Detailed Statistics and Collection of Dataset

Dataset coverage. Table 5 summarizes the dataset's 12 city settings and 22 study-region cases, with observations spanning 2015–2024. Cities includes Atlanta, Detroit, Mumbai, Phoenix, Beijing, Delhi, Guangzhou, Kunming, Łódź, Shanghai, Valencia and Seoul. The dataset contains 7,071 data

Table 5: Statistics and data-family composition of the multisource urban case dataset.

Aspect	**Count**	**Coverage**
Spatial and temporal coverage	***Count***	***Range***
Cities / Regions	**12 / 22**	City / study-region cases
Observation Years	**10**	2015–2024
Data composition	***No. of types***	***Included data***
Source Families	**8**	OSM; WorldPop; GHSL; VIIRS; Landsat; Sentinel-2; Mapillary; WorldCover
Map	**2**	Roads; POIs/AOIs
Indicator	**5**	Population; built-up areas; nighttime lights; Land Surface Temperature(LST); land cover
Overhead	**1**	Satellite imagery
Street	**1**	Street-view imagery
Modalities	**5**	Remote sensing; street view; raster; vector; text
Total Data Assets	**7,071**	-
Total Data Records	**1,507,693**	-

Table 6: Data sources, contents, analysis scales, and spatial resolutions of the multimodal urban case dataset.

Source	**Data Content and Benchmark Use**	**Analysis Scale**	**Resolution**	**Access**
OpenStreetMap	Administrative boundaries and vector features, including roads, buildings, land use, and amenities, for spatial delineation and accessibility analysis.	City/Region	–	OSM
WorldPop	Annual gridded population estimates for population change, density, and population-weighted accessibility analysis.	City/Region	1 km	WorldPop
VIIRS Annual VNL	Annual nighttime-light composites used as proxies for urban activity and its temporal changes.	City	15 arcsec	EOG
Landsat Collection 2 Level-2	Surface-temperature rasters for thermal analysis, with Landsat 8 RGB imagery supplementing early periods without available Sentinel-2 imagery.	City/Region	30 m grid	PC catalog
Sentinel-2 L2A	True-color satellite imagery for visual analysis of urban expansion, greenery, and surface change.	City/Region	10 m (RGB)	PC catalog
GHSL GHS-BUILT-S	Multi-epoch built-up-surface estimates for analyzing urban expansion, sprawl, and stagnation.	City	100 m	GHSL
ESA WorldCover	Land-cover maps distinguishing tree cover, grassland, cropland, built-up areas, water, and other classes.	City/Region	10 m	WorldCover
Mapillary	Geotagged street-view images for inspecting street environments, facilities, building conditions, and visible greenery.	Region	–	API docs

Note. Analysis scale denotes use within DUDA-Bench. Raster resolutions refer to the selected source products before benchmark-specific processing. Landsat 8/9 thermal observations have a native resolution of 100 m and are resampled to a 30 m product grid. Dashes indicate vector features or street-view photographs without a uniform ground sampling distance. PC denotes Microsoft Planetary Computer.

assets and 1,507,693 records across five modalities. Table 6 details each source's analytical use, spatial level, temporal coverage, resolution, and access point.

Table 7: Checkpoint rubric for trajectory-level evaluation.

ID	Checkpoint	Evaluation Criterion
C1	Issue Context	Identifies the urban issue and establishes an appropriate analytical context.
C2	Spatiotemporal Representation	Establishes the relevant spatial scope, temporal scope, and analysis units.
C3	Multimodal Evidence Planning	Selects or plans evidence sources that are relevant to the analytical objective.
C4	Indicator Acquisition	Obtains or derives the indicators needed to characterize the urban condition.
C5	Comparative Analysis	Performs appropriate temporal, spatial, or reference-based comparisons using the obtained evidence.
C6	Multimodal Evidence Integration	Relates evidence across relevant modalities rather than treating observations in isolation.
C7	Diagnostic Interpretation	Forms a diagnosis that is explicitly supported by the preceding observations and analyses.
C8	Problem Localization	Localizes the diagnosed phenomenon to appropriate spatial areas when required by the task.
C9	Mechanism Driver Analysis	Identifies plausible explanatory drivers grounded in observable evidence.
C10	Mechanism Chain	Connects supported drivers and observed urban outcomes into a coherent diagnostic chain without unsupported causal claims.
C11	Heterogeneity / Spillover	Examines relevant spatial heterogeneity, local variation, or spillover effects when applicable.
C12	Validation and Uncertainty	Checks intermediate conclusions against available evidence and acknowledges relevant uncertainty or limitations.
C13	Planning-Oriented Synthesis	Integrates the diagnosis and supporting evidence into planning-relevant conclusions or recommendations.
C14	Data Validity Awareness	Recognizes material data-quality, coverage, temporal, or modality limitations that affect the analysis.

Scoring: Pass (1.0): substantively completed and grounded in the trajectory; Partial (0.5): partially completed but incomplete or weakly grounded; Fail (0.0): absent, incorrect, or unsupported.

Table 8: Agreement between human verification and the LLM-as-a-Judge (GLM-5.3) on trajectory checkpoints. Three-level labels (*pass* / *partial* / *fail*); 716 checkpoint verdicts from 56 trajectories.

Evaluation	**Raw Agreement**	**Accuracy vs. Gold**
Annotator A vs. Annotator B	93.6%	–
Annotator A vs. GLM-5.3	89.8%	–
Human annotator (A)	–	97.8%
GLM-5.3	–	90.4%
Accuracy gap	–	7.4 pp

A.3 Extended evaluation details and Human Verification

Trajectory Evaluation Details. Table 7 defines the 14 checkpoints used by both the LLM-as-a-Judge and human annotators to evaluate agent trajectories. Each applicable checkpoint is scored as Pass (1.0), Partial (0.5), or Fail (0.0) based on its completion and support from the recorded trajectory.

Human Verification. We randomly sampled 56 ReAct workflow trajectories, with eight trajectories from each of the seven backbone models. Each trajectory was converted into an annotation packet containing exactly the information provided to the checkpoint judge: the task specification, checkpoint definitions with reference content, the compacted observable trajectory, and the final output, while the judge verdicts were withheld. Annotators A and B independently labelled every applicable checkpoint as *pass*, *partial*, or *fail* using the same written guideline, which restates the

Table 9: Domain skills and representative executable operations in DUDA-Env.

Capability	Domain Skill	Representative Operations	Function and Output
Data discovery	`data-card`	`aoi_data_inventory`; `data_access_guide`	Identifies available data sources, years, files, spatial scopes, and access methods for the current case.
Numerical analysis	`urban-metrics`	`annual_mean_metric`; `annual_total_metric`; `temporal_change_rate`; `temporal_series`; `landcover_class_share`; `population_weighted_facility_distance`; `raster_percentile`	Computes annual statistics, temporal changes, land-cover shares, accessibility measures, and raster percentiles.
Map and facility queries	`osm-map-query`	`poi_query`; `building_query`; `road_query`; `landuse_query`	Retrieves OSM roads, buildings, land-use polygons, and facilities such as hospitals, schools, and parks.
Satellite imagery access	`raster-imagery`	`temporal_satellite_retrieval`; `regional_satellite_retrieval`	Selects Sentinel-2 imagery by year and region, using Landsat imagery when the required early observation is unavailable.
Raster visualization	`raster-imagery`	`raw_raster_visualization`; `normalized_raster_visualization`; `thematic_raster_visualization`	Converts LST, population, built-up, nighttime-light, and land-cover rasters into interpretable maps and summary statistics.
Spatial localization	`spatial-localization`	`aoi_grid_generation`; `region_bbox_resolution`	Divides a city or study region into analysis units and resolves benchmark region identifiers to predefined spatial bounds.
Street-view inspection	`street-view`	`mapillary_streetview_sampling`	Samples geotagged Mapillary images and returns image files, locations, and capture metadata.
Custom geospatial analysis	`geospatial-sandbox`	`custom_geospatial_computation`	Executes Python for raster statistics, spatial joins, aggregation, visualization, and other task-specific operations.

judge rubric and specifies additional operational rules. Both annotators were blind to each other's annotations and to the judge outputs. Disagreements were resolved by a third reviewer applying the same guideline, producing the adjudicated human gold standard. As shown in Table 8, A and B agreed on 93.6% of the 716 checkpoint verdicts, corresponding to a 6.4% disagreement rate. Annotator A and GLM-5.3 agreed on 89.8% of checkpoint labels. Against the adjudicated gold standard, Annotator A reached 97.8% accuracy, while GLM-5.3 achieved 90.4%, a gap of 7.4 percentage points. When *partial* and *fail* are merged into a single non-pass class, the judge accuracy increases to 94.1%, indicating that most remaining errors arise from distinguishing partial completion from complete failure rather than from incorrectly identifying successful checkpoints.

A.4 Domain Skills

DUDA-Env provides domain skills that guide agents in discovering case data, computing urban indicators, querying map features, inspecting imagery, and localizing spatial patterns. Table 9 summarizes these skills alongside representative executable operations. A geospatial sandbox additionally supports task-specific analyses beyond the predefined operations.

B Extended Related Work

Data-Driven Urban Diagnosis. Urban problem diagnosis is an established research area in urban studies. Broad reviews characterize diagnosis as a basis for understanding urban conditions and informing appropriate interventions through historical, conceptual, and methodological perspectives Gu et al. (2018); Bibri (2021); Foster (2016). Empirical research further operationalizes this perspective through data-driven analyses tailored to individual urban problems and regions. Representative examples include urban sprawl and land fragmentation in Phoenix Shrestha et al. (2012), urban shrinkage and socio-spatial inequality in Detroit Xie et al. (2018), surface urban heat island patterns in Valencia Wei & Sobrino (2024), and accessibility in Seoul Kang & Lee (2025). Together, these studies demonstrate how heterogeneous spatial, temporal, and socioeconomic evidence can support urban diagnosis. However, such analyses are generally formulated around individual problem settings rather than as a shared computational task spanning multiple urban issues, regions, and

data sources. DUDA-Bench builds on this literature by abstracting their common analytical structure into a unified agent-oriented formulation.

Benchmarks for Urban and Geospatial Intelligence. Recent benchmarks have expanded urban intelligence evaluation along several complementary directions. Interactive benchmarks evaluate agents operating within urban environments: CityBench combines simulator-based understanding and decision making Feng et al. (2025b), CityEQA studies embodied question answering through city exploration Zhao et al. (2025), and USTBench evaluates spatiotemporal reasoning at both process and outcome levels Lai et al. (2026). A second line focuses on multimodal urban understanding. UrBench evaluates satellite, street-view, and cross-view reasoning Zhou et al. (2025), UBench integrates spatial, temporal, and visual urban information Feng et al. (2025a), DynamicVL targets multi-temporal urban change Xuan et al. (2026), and CityLens connects urban imagery with socioeconomic indicators Liu et al. (2026b). Related geospatial and Earth-observation benchmarks extend this scope with multi-sensor vision Danish et al. (2025), agentic EO workflows Feng et al. (2026), and executable GIS analysis Yu et al. (2026). These benchmarks provide rich urban and geospatial evaluation settings, but their primary targets are individual perception, reasoning, prediction, navigation, or GIS capabilities. DUDA-Bench instead treats urban problem diagnosis itself as the evaluation target and requires multiple capabilities to be composed within a shared multi-stage workflow.

LLM Agent Benchmarks. LLM agent evaluation has increasingly shifted from final-answer accuracy toward interactive execution and trajectory-level reliability. In general settings, τ-bench evaluates consistent success in tool-mediated interactions using pass^k Yao et al. (2025); GTA-2 connects atomic tool use with open-ended workflows through checkpoint-based evaluation Wang et al. (2026); and TRAJECT-Bench evaluates tool selection, parameter grounding, and ordering along execution trajectories He et al. (2026). Scientific-agent benchmarks pursue a similar direction: SciAgentBench spans elementary actions and multi-tool workflows Shen et al. (2026), MoSciBench evaluates coding agents for multimodal data-driven discovery Liu et al. (2026a), and ZephyrusBench grounds agent evaluation in programmatic weather analysis Varambally et al. (2026). DUDA-Bench follows this process-aware evaluation paradigm while extending it to multimodal urban diagnosis, where agents must connect problem framing, heterogeneous evidence acquisition, diagnostic reasoning, and validation within an end-to-end analytical process.

C Agent System Environment Configuration

We instantiate five representative agent systems under a shared benchmark environment. All systems receive the same task input, public data card, benchmark-provided data and tool interfaces, and evidence-grounding policy, while preserving their system-specific execution and orchestration mechanisms. Unless otherwise specified, agents are restricted to the provided case bundle and cannot access external web resources.

C.1 ReAct System Prompt

ReAct System Prompt

Role. You are an expert urban-analysis agent. You answer questions about a city or region by inspecting multimodal geospatial data and writing analysis code.

Output Protocol.
At every step, output a SINGLE JSON object and no text outside the JSON. Choose exactly one action:

```
{"thought": "...", "action": "run_code",
 "code": "..."}
```

```
{"thought": "...", "action": "call_tool",
 "tool_name": "...", "tool_args": {...}}
```

```
{"thought": "...", "action": "finish",
 "final_answer": "..."}
```

Execute at most one environment action at each interaction step. After receiving the corresponding observation, decide the next action based on the updated evidence.

Constraints.
1. Use only the provided benchmark data and environment. Do not browse the web or access files outside the supplied case bundle.
2. Ground substantive findings in observable evidence. Identify the corresponding data source, modality, artifact, or path when available.
3. If the available evidence is insufficient, state the limitation explicitly instead of guessing.
4. The complete current AOI data card is already provided below. Inspect it directly and do not invoke tools merely to rediscover the same available modalities.
5. There is no fixed priority among numeric, visual, textual, and code-based evidence. Select actions according to the task and the available modalities.
6. Qualitative judgments directly observable from RGB imagery (e.g., denser built-up areas, greener land cover, or visible land-cover change) are valid visual evidence. Thermal conditions and exact quantitative indices must be derived from the corresponding canonical raster, summary, or analytical data rather than inferred visually.
7. Do not inspect Python source code, module dictionaries, configuration files, settings, or unrelated directory listings as task evidence.
8. Follow the exact output format requested by the task.

Execution Support.
The agent additionally receives the benchmark sandbox API description used for code execution. The full configuration also exposes the benchmark-provided domain-skill instructions and available skill directories. These components constitute the full ReAct system evaluated in the main experiments.

Public Data Card.

```
<public_data_card>
{public_data_card}
</public_data_card>
```

C.2 Direct Tool Use System Prompt

Direct Tool Use System Prompt

Role. You are an expert urban-analysis assistant. Solve the task directly, without using a fixed multi-step agent framework, iterative think–act–observe loop, or explicit planning stage. You may use the provided environment actions for one data-gathering round before answering, or answer immediately if the provided evidence already suffices.

Constraints.
1. You have at most ONE round of environment actions. All required actions must be selected before observing any tool result.
2. Tool results cannot be used to generate additional actions or modify the arguments of previously selected actions. After all selected actions are executed, you must produce the final answer.
3. Use only the provided benchmark data and environment. Do not browse the web or access files outside the supplied case bundle.
4. Ground substantive findings in observable evidence. Identify the corresponding data source, modality, artifact, or path when available.
5. If the available evidence is insufficient, state "insufficient evidence" instead of guessing.
6. Qualitative judgments directly observable from RGB imagery (e.g., denser built-up areas, greener land cover, or visible land-cover change) are valid visual evidence. Thermal conditions and exact quantitative indices must be derived from the corresponding canonical raster, summary, or analytical data rather than inferred visually.
7. The supplied public data card describes the available case data and should not be rediscovered through redundant environment calls.
8. Follow the exact output format requested by the task.

Output Protocol.
The FIRST reply must be a SINGLE JSON object and contain no text outside the JSON. Choose one of the following forms:

```
{"thought": "...",
 "tool_calls": [
   {"tool_name": "...", "tool_args": {...}},
   ...
 ]}
```

or

```
{"thought": "...",
 "final_answer": "..."}
```

Request all required actions in the single `tool_calls` list. The runtime may truncate the list according to the configured environment-action budget.

Public Data Card.

```
<public_data_card>
{public_data_card}
</public_data_card>
```

C.3 PLAN-AND-EXECUTE SYSTEM PROMPT

Plan-and-Execute System Prompt

Planning Role. You are an urban-analysis planning agent. Given a question and the available environment actions with their input schemas, produce a short, ordered plan whose execution gathers sufficient evidence to answer the question.
The complete plan must be finalized before any environment observation is seen. During execution, observations may be recorded for final synthesis but must not alter, reorder, add, or remove planned actions.

Planning Constraints.
1. Produce an ordered plan of environment actions, including tool calls or code execution when available.
2. You will not observe action results before finalizing the complete plan and cannot replan afterward.
3. Use only action names provided by the benchmark environment and supply all fields required by their input schemas.
4. Use only the provided benchmark data and environment. Do not browse the web or access files outside the supplied case bundle.
5. The public data card already describes the available case data; do not plan redundant actions merely to rediscover available modalities.
6. Use at most {`max_steps`} planned environment actions.

Planning Output Protocol.
Respond with a SINGLE JSON object and no text outside the JSON:

```
{
  "plan": [
    {
      "step": 1,
      "tool_name": "...",
      "tool_args": {...},
      "reason": "..."
    }
  ]
}
```

Synthesis Role.
You are an urban-analysis agent. The previously generated fixed plan has already been executed. Answer the original question using ONLY the evidence obtained from those actions.

Synthesis Constraints.
1. Ground substantive claims in the actually observed evidence and identify the corresponding data source, modality, artifact, or path when available.

2. If an action failed or the available evidence is insufficient, state the unresolved limitation rather than inventing additional evidence.
3. Do not request further environment actions or revise the original plan.
4. Follow the exact output format requested by the task.

Public Data Card.

```
<public_data_card>
{public_data_card}
</public_data_card>
```

C.4 MAF System Prompts

We instantiate the Microsoft Agent Framework (MAF) baseline using a manager–analyst–reviewer organization. All three roles use the same backbone model. Only the Analyst is allowed to execute benchmark environment actions; the Reviewer performs read-only verification, while the Manager coordinates the workflow and produces the final response.

MAF Analyst, Reviewer, and Manager Prompts

Common Policy.
Use only the provided benchmark data and environment. Do not browse the web or access files outside the supplied case bundle. Ground substantive findings in observable evidence and identify the corresponding data source, modality, artifact, or path when available. If evidence is insufficient, state the limitation rather than guessing. Follow the exact output format requested by the task.

Analyst.
You are the urban-data Analyst and the only role allowed to perform analytical environment actions.
You may invoke `dura_run_code` and `dura_call_tool`. Select the appropriate interface according to the analytical need and available evidence. Inspect returned observations and revise or debug the analysis when necessary. Use environment actions efficiently and avoid redundant calls.
Do not inspect Python source code, module dictionaries, settings, configuration files, or unrelated directory listings; these are not task evidence. Once the assigned evidence has been collected, return the evidence and any unresolved limitations to the Manager rather than continuing environment discovery.
Across the entire case, code execution and domain or visual tools share a budget of {`max_environment_actions`} executed environment actions. If a tool reports an `ANALYST_*_BUDGET_EXHAUSTED` condition, immediately return the best available evidence and unresolved limitations and do not retry the exhausted action.
Available DURA tool schemas are provided by the runtime.

Reviewer.
You are a read-only evidence and answer Reviewer and have no environment execution tools.
Review the Analyst's evidence, calculations, citations, and coverage. Identify specific substantive gaps, unsupported claims, or analytical inconsistencies. Do not invent observations or request purely cosmetic revisions.
If the evidence and current answer substantively satisfy the original request, begin with `APPROVED`. Otherwise, begin with `NEEDS_REVISION` and list only blocking corrections.

Manager.
You are the Magentic task Manager. Coordinate the Analyst and Reviewer and produce the final answer.
Plan and coordinate the participants, but do not perform analysis yourself and do not claim observations that the participants did not produce. Treat internal planning and progress-ledger requests as orchestration protocol messages rather than benchmark answers.
When the benchmark requests JSON, return one complete and compact JSON object with no markdown or additional text.

Public Data Card.

```
<public_data_card>
{public_data_card}
</public_data_card>
```

C.5 LAMBDA SYSTEM PROMPT

For LAMBDA, we use a lightweight benchmark-facing adaptation rather than its original interactive interface. Following the sequential tool-execution setup used for data-science-agent evaluation in EvoDS Yang et al. (2026), the agent iteratively selects an analytical function, observes its actual execution result, and adapts subsequent actions. This preserves LAMBDA's data-analysis-oriented iterative execution while replacing external execution components with the common DURA environment for controlled comparison.

LAMBDA System Prompt

Role.
You are a LAMBDA-style data-analysis agent operating on a DURA-Bench case. Solve the task by repeatedly selecting a native analytical function, inspecting its real observation, and adapting the next step. When sufficient evidence has been collected, return the final answer as ordinary text and make no further function call.

Execution Policy.
1. Use only the supplied functions and public AOI bundle. Never browse the web, invoke an external shell, access host files, or invent observations.
2. Use environment actions efficiently and avoid redundant calls. Do not repeat an unchanged failed or successful action. The hard limit is {`max_environment_actions`} environment actions.
3. `execute_python` provides a persistent analysis sandbox: serializable variables and data frames may persist across execution cells, and common imports are refreshed automatically when needed.
4. Inspect returned observations before deciding the next analytical action and revise the analysis when execution results indicate an error or inconsistency.
5. Ground substantive findings in observable evidence and identify the data paths, modalities, and years actually inspected when available.
6. If the available evidence is insufficient, state the limitation explicitly rather than guessing.
7. Visual tool outputs may contain actual image pixels. Use them for qualitative judgments directly supported by the image, but do not infer exact quantitative raster values or thermal conditions merely by visual inspection.
8. Do not inspect source code, configuration files, module dictionaries, or unrelated directory listings as analytical evidence.
9. Follow the exact output format requested by the task.

Workflow Protocol.
The runtime supplies the benchmark-specific workflow protocol describing the available native functions, their schemas, execution rules, and termination conditions.

Public Data Card.

```
<public_data_card>
{public_data_card}
</public_data_card>
```