UrAgent: Structuring Agentic Automation for Transparency and Agency in Urban Governance

ZHAOFENG NIU, Qufu Normal University, China
YUXIAO LIU, Qufu Normal University, China
ZHOUQIANG JIANG, Osaka University, Japan
BOWEN WANG, Osaka University, Japan
ISIDRO BUTASLAC, Nara Institute of Science and Technology, Japan
YIMING SHEN, Xi'an University of Posts and Telecommunications, China
LIANGZHI LI, Qufu Normal University, China

As AI systems increasingly integrate perception, reasoning, and action within unified interfaces, understanding their impact on organizational oversight becomes critical. In urban governance, multimodal workflows rely on spatial data that must remain accountable and auditable. We introduce **UrAgent**, a multimodal agent that structures urban reasoning into staged interpretation and execution. By separating contextual analysis from tool-level operations and externalizing execution decisions, the system makes intermediate steps visible and traceable. Experiments on the UrBench benchmark show improved performance in geo-localization, scene reasoning, and object grounding, with reduced spatial inconsistencies compared to end-to-end baselines. Beyond accuracy, UrAgent enables clearer error localization and post-hoc inspection of decision pathways. These results demonstrate how architectural design shapes oversight, responsibility allocation, and long-term human–AI collaboration, highlighting the importance of inspectable coordination in high-stakes domains.

CCS Concepts: • **Human-centered computing → HCI design and evaluation methods**; **Interaction paradigms**; • **Computing methodologies → Agents and autonomous systems**.

Additional Key Words and Phrases: Multimodal Agent, Urban Governance, Multimodal Large Language Models

ACM Reference Format:
Zhaofeng Niu, Yuxiao Liu, Zhouqiang Jiang, Bowen Wang, Isidro Butaslac, Yiming Shen, and Liangzhi Li. 2026. UrAgent: Structuring Agentic Automation for Transparency and Agency in Urban Governance. In *Proceedings of AutomationXP26 Workshop of the 2026 CHI Conference on Human Factors in Computing Systems, April 14, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 6 pages.

1 Introduction

Agentic AI systems are increasingly deployed in organizational environments, where they operate not as isolated tools but as coordinated systems capable of initiating actions and invoking external resources [13, 15]. As automation shifts from discrete tools to multi-component agentic systems, the experience of oversight and responsibility in organizations

Authors' Contact Information: Zhaofeng Niu, Qufu Normal University, Rizhao, China, zhaofengniu@qfnu.edu.cn; Yuxiao Liu, Qufu Normal University, Rizhao, China, LYX_1126@qfnu.edu.cn; Zhouqiang Jiang, Osaka University, Osaka, Japan, zhouqiang@is.ids.osaka-u.ac.jp; Bowen Wang, Osaka University, Osaka, Japan, wang@ids.osaka-u.ac.jp; Isidro Butaslac, Nara Institute of Science and Technology, Nara, Japan, isidro.b@naist.ac.jp; Yiming Shen, Xi'an University of Posts and Telecommunications, Xi'an, China, doubaohero@aliyun.com; Liangzhi Li, Qufu Normal University, Rizhao, China, conscienceli@gmail.com.

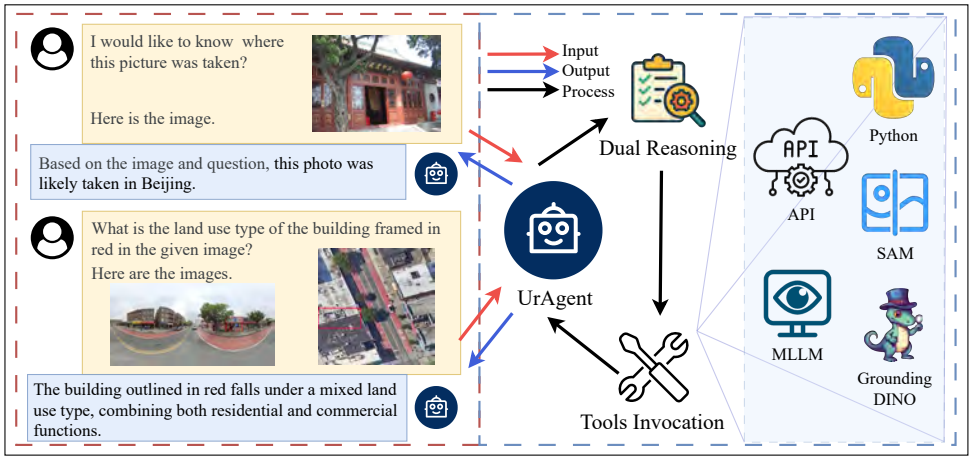

Fig. 1. Overview of UrAgent architecture with dual reasoning and tool invocation.

is increasingly reshaped. When multiple computational components jointly produce decisions, the locus of responsibility and the visibility of reasoning processes become less clear and more difficult to interpret.

These challenges are especially salient in urban governance. Municipal workflows rely on multimodal data—such as satellite imagery, street views, and textual reports—under conditions requiring accountability and reproducibility in practice [7]. End-to-end multimodal large language models (MLLMs) achieve strong performance, yet they implicitly merge perception, reasoning, and response generation into a single inference step [10]. As a result, spatial inconsistencies or hallucinated attributes are difficult to attribute to specific processes, complicating oversight, calibrated reliance, and the broader automation experience of human operators in real-world settings [5, 8].

We present UrAgent, a multimodal system based on a staged architecture that separates environment reasoning from tool reasoning and externalizes execution decisions in practice. Rather than optimizing solely for predictive accuracy, the design introduces explicit attribution boundaries between interpretation and action. We evaluate UrAgent on UrBench, an existing benchmark for multimodal urban reasoning tasks. Results show consistent gains in geo-localization and scene reasoning tasks. More importantly, the staged design exposes execution traces that support error localization and post-hoc inspection in complex scenarios.

The main contributions can be summarized as follows:

- We introduce a staged multimodal architecture that separates perception, reasoning, and action selection for governance-oriented applications in complex urban settings.
- We empirically demonstrate that architectural staging improves robustness and reduces spatial failure modes on UrBench. Through controlled comparisons, we show that decomposing reasoning and tool invocation leads to more stable performance across geo-localization and scene reasoning tasks.
- We analyze how attribution boundaries influence oversight capacity and long-term human–AI collaboration in broader organizational contexts.

2 Structuring Agentic Automation in Urban Governance

As shown in Figure 1, UrAgent operationalizes a multimodal large language model as a controller within a structured reasoning-and-execution loop. The architecture consists of three stages: dual reasoning, tool invocation, and structured response generation. The central design goal is not merely performance improvement, but the creation of explicit attribution points within the automation process.

2.1 Dual Reasoning as Attribution Boundary

Given multimodal inputs, UrAgent first performs environment reasoning. In this stage, the model parses visual and textual context to identify relevant environmental attributes, spatial relationships, and ambiguities. This step produces an intermediate structured representation that externalizes interpretation before any execution decision is made. Next, tool reasoning determines whether external tools should be invoked. By separating contextual interpretation from execution planning, UrAgent creates a responsibility boundary between perception and action. This decomposition reduces the conflation of semantic inference with procedural decision-making. From an automation experience perspective, dual reasoning introduces inspectable checkpoints. Human operators can examine whether errors originate from misinterpretation of environmental context or inappropriate tool selection. Such separation supports traceability and calibrated reliance.

2.2 Rule-Based and Inspectable Tool Invocation

Following dual reasoning, UrAgent maps the structured representation to a tool execution plan. Unlike end-to-end implicit coordination within foundation models, tool scheduling in UrAgent follows explicit rule-based mappings between task types and candidate tools. Each tool invocation forms an accountability unit. For example, segmentation tools handle instance-level perception, while detection modules support relational reasoning. If no external computation is required, the controller defaults to internal reasoning. This design transforms implicit agent coordination into explicit, inspectable coordination. Rather than obscuring which sub-process contributed to an outcome, UrAgent exposes the sequence of reasoning, selection, and execution steps.

2.3 Structured Response Generation

In the final stage, tool outputs are integrated with multimodal inputs to produce a coherent response. Importantly, responses can be traced back to specific reasoning and execution steps. This enables post-hoc inspection of whether an answer relies on internal inference, external perception tools, or a combination of both. By preserving intermediate representations and tool outputs, UrAgent supports retrospective auditing and reproducibility—properties that are essential in governance-oriented automation systems.

3 Empirical Evaluation

We evaluate UrAgent on UrBench [16], an established benchmark for multimodal urban reasoning. Experimental settings remain identical to baseline MLLMs to isolate architectural effects. Table 1 shows that UrAgent improves performance across geo-localization, scene reasoning, and object grounding tasks. Gains are particularly evident in counting, comparison, and spatial localization tasks. We observe fewer spatial hallucinations and cross-view inconsistencies compared to end-to-end baselines, suggesting that staging contributes to improved failure localization. Controlled comparisons further indicate that removing tool invocation reduces grounding accuracy, while eliminating environment

Table 1. Results on Geo-Localization and Scene Understanding tasks with overall average scores. Bold indicates the highest, underline indicates the second-highest, ⋆ indicates no external tools are used.

Model	Geo-Localization				Scene Understanding				Scene Reasoning			Object Understanding		
	CR	IR	CL	OR	SR	RU	CO	SC	RBR	TSR	VPR	OM	OG	OAR
Human	30.0	92.6	82.9	85.7	59.2	87.2	94.1	85.1	87.4	85.7	88.2	95.2	95.5	61.6
Random	24.8	23.9	25.1	23.2	17.7	25.7	21.4	25.3	23.9	24.2	30.6	21.8	22.1	21.5
GPT-4o [1]	79.2	85.9	35.3	30.7	65.0	66.3	40.1	79.0	**79.6**	68.2	77.9	28.0	46.5	50.1
Gemini-1.5-Flash [14]	69.7	25.9	25.9	24.0	57.9	**71.0**	29.1	67.7	77.8	**75.8**	69.8	22.0	39.1	40.9
LLaVA-NeXT-7B-Vicuna [10]	51.2	24.6	27.2	23.3	56.1	20.2	34.3	25.9	49.6	51.5	54.5	31.8	27.2	31.9
Mantis-LLAMA3-SigLIP [6]	67.0	32.4	27.0	27.2	59.2	44.5	27.4	52.4	67.6	41.6	57.7	25.2	34.2	38.6
InternVL2-8B [4]	50.8	26.6	31.8	25.2	53.0	52.6	43.0	51.4	74.9	54.8	62.6	30.3	32.0	41.3
LLaVA-NeXT-13B [10]	52.0	24.5	27.7	26.7	53.9	50.7	33.8	25.1	54.0	52.1	52.3	31.8	34.8	26.3
VILA-1.5-13B [9]	62.7	33.7	28.6	24.1	47.7	43.9	23.9	48.6	66.3	43.8	46.4	25.8	32.3	38.2
InternVL2-26B [4]	61.3	23.0	32.3	24.7	65.0	41.1	30.1	52.6	77.9	63.8	71.2	26.1	37.3	48.4
LLaVA-NeXT-34B [10]	58.4	26.0	28.5	27.8	58.3	49.0	21.6	53.9	65.6	59.3	56.8	28.7	40.5	24.1
VILA-1.5-40B [9]	70.1	62.5	36.8	27.9	53.6	51.7	32.1	66.7	76.4	55.5	61.3	34.1	48.3	39.5
Ours(GPT-4o)	⋆**79.5**	**93.7**	**53.9**	**46.3**	⋆**72.1**	⋆64.9	**67.2**	**83.1**	77.3	69.0	**81.6**	**56.1**	**78.1**	**63.4**
Ours(GPT-4o mini)	⋆71.3	86.3	47.1	42.3	⋆63.2	⋆61.4	58.3	77.6	74.3	64.2	73.9	44.2	65.2	57.1

reasoning decreases cross-view localization performance (see Figure 2). Beyond accuracy, the staged design enables clear error attribution. Incorrect outputs can be traced to either contextual interpretation or tool limitations, improving failure visibility in multi-component workflows.

4 Implications for Automation Experience

UrAgent serves as a case study for how architectural design shapes automation experience in organizations. Beyond model accuracy, it shows how decomposition influences transparency, responsibility, and long-term human–AI collaboration. This concern resonates with levels-of-automation theory, which argues that increasing autonomy without transparency can undermine operator control and situational awareness [12]. It also aligns with socio-technical perspectives emphasizing that technical architectures reshape organizational responsibility structures [3].

4.1 Architectural Attribution in Multi-Component Systems

A core challenge in agentic automation is attribution: when multiple components contribute to an outcome, responsibility becomes unclear. In governance workflows—such as infrastructure inspection or environmental assessment—decisions must remain auditable and reproducible. By separating environment reasoning from tool reasoning, UrAgent establishes explicit attribution boundaries between contextual interpretation and procedural execution. Each tool invocation becomes an accountability unit rather than an implicit computation. This design aligns with calls for inspectable AI coordination over opaque end-to-end inference [2]. Rather than exposing raw internal states, it organizes visibility around key decision checkpoints, reducing ambiguity in multi-component automation.

4.2 Calibrated Human Agency in Agentic Workflows

Automation research highlights risks of over-reliance and automation bias when systems lack interpretability [8]. In urban governance, where decisions affect public resources, maintaining meaningful human agency is essential. UrAgent introduces intervention points within the reasoning pipeline. Operators can inspect intermediate representations,

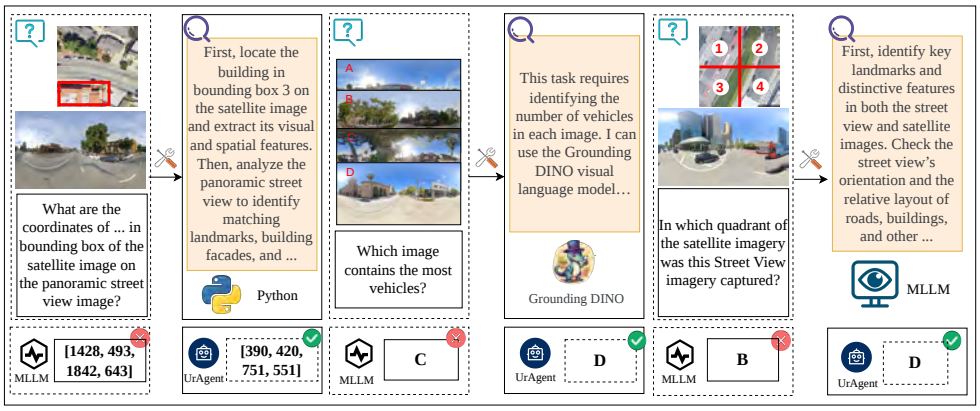

Fig. 2. Comparison between an end-to-end multimodal baseline and UrAgent.

reassess tool selections, or re-execute perception stages when outputs conflict with domain knowledge. This shifts automation from full delegation to calibrated delegation, supporting appropriate reliance instead of blind trust [5]. Architectural structuring thus distributes authority while preserving oversight.

4.3 Sustaining Expertise in Long-Term Human–AI Collaboration

Long-term use of agentic systems raises concerns about deskilling. When reasoning collapses into opaque outputs, operators may disengage from interpretive processes. By externalizing intermediate reasoning and preserving spatial inference steps, UrAgent promotes co-reasoning rather than cognitive offloading. Analysts remain engaged in interpreting environmental attributes and relational constraints, maintaining domain understanding while benefiting from computational support. Designing agentic systems that preserve interpretive visibility may therefore support sustainable organizational integration of AI technologies, echoing human-centered design principles that emphasize visibility of system state [11].

5 Conclusion

This paper presents UrAgent as a structured approach to agentic automation in urban governance. Beyond performance gains, it shows how reasoning decomposition and tool-level attribution support transparency and calibrated human agency. By reframing multimodal agents as structured pipelines rather than monolithic predictors, this work contributes to discussions on automation experience, accountability in multi-component AI systems, and sustainable human–AI collaboration. It highlights that architectural structuring is not only a technical choice but also a design intervention shaping how humans experience and govern agentic AI systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62372266), Shandong Provincial Natural Science Foundation for Excellent Young Scholars (2024HWYQ-075), Shaanxi Provincial Department of Education (25JS114), Taishan Scholar Program (tsqn20221133), Rizhao Science Fund (RZ2022ZR01), and the Rizhao–Qufu Normal University Joint Technology Transfer Center.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[3] Gordon Baxter and Ian Sommerville. 2011. Socio-technical systems: From design methods to systems engineering. *Interacting with computers* 23, 1 (2011), 4–17.

[4] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences* 67, 12 (2024), 220101.

[5] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[6] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483* (2024).

[7] Rob Kitchin. 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

[8] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[9] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 26689–26699.

[10] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[11] Don Norman. 2013. *The Design of Everyday Things*. Basic Books.

[12] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30, 3 (2000), 286–297.

[13] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[14] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[15] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First conference on language modeling*.

[16] Baichuan Zhou, Haote Yang, Dairong Chen, Junyan Ye, Tianyi Bai, Jinhua Yu, Songyang Zhang, Dahua Lin, Conghui He, and Weijia Li. 2025. Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10707–10715.