

CityReal: Human-Aligned Urban Behavior and City Dynamics Simulation with Large-Scale LLM Agents

Nicolas Bougie¹, Xiaotong Ye¹, Narimasa Watanabe¹

{nicolas.bougie, tony.yip, narimasa.watanabe}@woven.toyota

¹Woven by Toyota

Abstract

Large-scale urban simulation plays a pivotal role in social science, traffic safety, and transportation policy. Recent work has shown that large language models, when prompted as agents, can generate lifelike daily routines at city scale. Yet these methods typically rely on few-shot prompting, causing agents to reproduce the LLM’s behavioral priors rather than the target population. We introduce CityReal, a modular framework for human-aligned urban simulation. CITYREAL models agents as intention-driven decision makers that pursue coherent mobility and activity plans rather than isolated step-by-step choices. They adapt over time by learning habits and preferences based on experience and constraints. To improve population-level realism, we learn textual adapters for behavior modules that align agent decisions with observed population statistics. Experiments show that CITYREAL improves alignment with real-world human behavior at both micro and macro levels. Scaling to tens of thousands of agents, it supports analysis of crowd density, place popularity, mobility flows, and well-being under different urban scenarios, offering a scalable testbed for urban simulation and forecasting.

1 Introduction

Reproducing the daily behavior of urban populations is a long-standing challenge in computational social science (Lazer et al., 2009; Hofman et al., 2021), with applications ranging from traffic safety, transportation analysis to urban planning. Traditional simulators typically model individuals through hand-crafted behavioral rules (Epstein, 1999; Macal and North, 2005). While transparent and scalable, these rules can make behavior rigid and difficult to adapt to unseen situations (Zheng et al., 2022; Wang et al., 2023a; Feng et al., 2025). Thus, capturing the psychological, social, and environmental factors that shape urban behavior remains challenging.

The emergence of large language models (LLMs) offers a new direction. Rather than relying on fixed decision rules, LLM-based agents can reason in natural language and choose actions conditioned on their demographic profile, internal state, and surrounding environment (Park et al., 2023a; Gao et al., 2024). Recent work has applied this paradigm to urban simulation. LLMob (Wang et al., 2024) generates resident trajectories. AgentSociety (Piao et al., 2025) studies collective social phenomena such as opinion polarization and policy response. CitySim (Bougie and Watanabe, 2025a) and MobileCity (Ye et al., 2026) scale LLM-agent simulation to large populations with richer personas, memory, and internal needs.

Despite this progress, existing LLM-based urban simulators face several limitations. Prior approaches often rely on few-shot prompting to imitate plausible residents (Wang et al., 2024; Bougie and Watanabe, 2025a), without explicitly aligning agent decisions with observed human behavior. Besides, agents often make discretionary decisions as isolated choices, producing trajectories with limited continuity, whereas real behavior is organized around persistent intentions such as running errands or spending time in a particular area. Finally, agents rarely *learn* from their own history. Their behavior is typically determined by the initial persona and prompts, limiting the emergence of habits, preferences, social tendencies, and practical constraints accumulated through experience. As a result, agents with similar demographic profiles may behave too similarly instead of developing distinct routines shaped by their own trajectories.

We propose CityReal, an LLM-based framework for simulating urban populations whose behavior is both coherent at the individual level and aligned with human patterns. First, we learn textual adapters with Monte Carlo Tree Search to calibrate agent decisions toward target population statistics while keeping the LLM frozen. Second,

agents organize behavior around persistent intentions, maintaining stable motives across consecutive actions rather than selecting each activity independently. Third, CityReal introduces experience-driven adaptation through end-of-day reflection, allowing agents to gradually accumulate habits, preferences, social tendencies, and constraints from their own experience. Together with persona, memory, belief, mobility, and social modules, these mechanisms produce urban agents that can adapt to context while remaining calibrated to observed human behavior. This marks a paradigm shift in urban population simulation, enabling large-scale, human-aligned agent populations that capture the behavioral complexity and diversity of real-world societies.

2 Related Work

Replicating human behavior in urban environments remains a challenge (Hofman et al., 2021; Lazer et al., 2009). Traditional agent-based models have been widely used to study complex phenomena, resource allocation, and policy evaluation (Epstein, 1999; Macal and North, 2005; Wilensky and Rand, 2015). Yet, they depend on hard-coded rules or fixed utility functions, constraining their capacity to capture behavioral diversity, adaptability, and long-term dynamics (Feng et al., 2025; Zheng et al., 2022; Wang et al., 2023b). Frameworks such as CityBench (Feng et al., 2025) and AI4SIM (Zheng et al., 2022) have attempted to integrate richer, data-driven approaches, yet realistic cognitive and motivational modeling remains limited. Recently, LLMs have opened new possibilities for simulating human-like agents in virtual worlds (Park et al., 2023a; Gao et al., 2024; Li et al., 2023; Wei et al., 2022; Bougie et al., 2026c). LLM-powered agents can reason, plan, and interact through natural language (Gao et al., 2024; Wei et al., 2022; Bougie and Watanabe, 2025b; Park et al., 2023a; Bougie and Watanabe, 2024). Hong et al. (2024) demonstrates how agents can collaborate in complex software engineering tasks. Few studies have also explored agent alignment, including user simulation for recommender-system evaluation (Bougie et al., 2026a) and alignment to human interaction trajectories (Bougie et al., 2026b). However, these methods depend on human behavior data for each simulated agent, making them difficult to scale to urban populations. As research moves toward larger-scale simulations, computational efficiency becomes cru-

cial. Park et al. (2024) scale up simulations to 1,000 agents, but still inherits prohibitive costs. To improve realism, AgentSociety (Piao et al., 2025) and MobileCity (Ye et al., 2026) use episodic memory and gravity-based place selection. Similarly, CitySim (Bougie and Watanabe, 2025a) advances LLM-based urban simulation through recursive activity planning, dynamic memory, and belief modules. Yet existing approaches still struggle to align simulated populations with observed human behavior while preserving coherent and diverse individual trajectories. They also lack mechanisms for maintaining long-term behavioral coherence or allowing agents to develop habits and preferences from their own experiences.

3 Method

We propose **CityReal**, an LLM-based framework for simulating urban populations whose behavior is both individually coherent and aligned with observed real-world patterns (Figure 1). We model each resident as an agent embedded in a graph-structured city, where nodes correspond to places or areas and edges denote mobility connections (e.g., car, train). Each agent is initialized with a persona, after which CityReal calibrates behavior modules to target population statistics by learning textual adapters that condition their decisions. During simulation, an orchestrator selects the appropriate module from the agent’s current context. Agents follow a perception–decision–reflection loop: they perceive the environment, retrieve relevant memories, continue or revise their behavior, and update habits, preferences, and beliefs through end-of-day reflection.

3.1 Cognitive State Representation

Persona Module. The persona encodes an agent’s background: demographic attributes (age, gender, occupation, salary, household composition, life stage), spatial anchors (home, workplace, school), and psychographic traits (personality, lifestyle, routine tendencies). These attributes shape feasible activities and daily behavior (e.g., students follow school schedules, full-time workers commute on weekdays, and retirees maintain more flexible routines).

Memory Module. The memory module connects past experience to future behavior through three components. **Temporal memory** records experiences in chronological order, including time, loca-

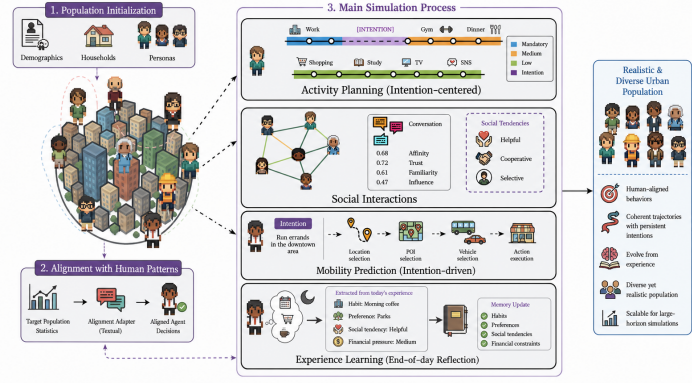


Figure 1: Overview of CityReal: LLM-based agents with diverse personas plan daily activities, interact socially, and navigate a virtual city environment.

tion, activity, observation, and outcome. **Spatial memory** stores beliefs about places, such as affordability, convenience, crowding, atmosphere, and enjoyment; beliefs are updated after visits, while unvisited POIs are initialized from similar known places. **Reflective memory** stores conclusions distilled from experience. Unlike temporal memory, which logs what happened, reflective memory captures what the agent has learned and is retrieved during later decisions.

Needs and Financial Module. Agents track short-term needs (hunger, energy, safety, social connection, financial security) that evolve throughout the day and may interrupt ongoing routines. In this study, we explicitly model financial constraints through each agent’s income, recent spending, and available budget, from which financial pressure is derived. This pressure directly shapes activity and destination choice, capturing socioeconomic heterogeneity.

Belief Module. Following CitySim (Bougie and Watanabe, 2025a), the belief module maintains and updates the agent’s subjective evaluations of places and social contacts. After each POI visit, the agent produces an assessment of the experience, covering affordability, convenience, crowding, and enjoyment, which is integrated into spatial memory. For social contacts, beliefs over affinity, trust, and familiarity are revised after interactions, shaping future decisions such as whom to contact or whether to participate in group activities.

3.2 Planning and Mobility

Daily schedules are generated via a recursive decomposition of time into [blocks], each including a starting time, duration, and activity or inten-

tion. At the beginning of the day, planning first assigns mandatory tasks (e.g., sleep, work), then recursively fills remaining [EMPTY] blocks with medium-priority tasks (e.g., meals, hygiene).

The remaining flexible blocks are not pre-selected, instead, each is instantiated at execution time through the intention module. When a flexible block begins, the agent forms an intention — a natural-language description of what it is trying to accomplish, including an optional area anchor, an expected duration, and a completion status. Examples include *run errands near the station* or *find an inexpensive meal before returning home*. The current intention provides a shared context for subsequent decisions, preventing flexible activities from being generated independently at each step. During execution, the orchestrator determines whether the active intention should be continued, revised, completed, or replaced.

3.2.1 Place Selection

When an activity requires movement, the agent selects a destination in two stages.

Area Selection. The agent first selects a broad area based on the current intention, location, schedule, beliefs, social context, and financial pressure.

POI Selection. Within the selected area, candidate POIs are evaluated using a belief-aware gravity model:

$$w_{i,j}^t = \frac{\exp(\beta^\top f_{i,j}^t)}{(1 + D_{i,j}^t)^\gamma}, \quad (1)$$

where $D_{i,j}^t$ is the distance from agent i to POI j , $f_{i,j}^t \in [0, 1]^d$ is a vector of structured belief features (price, crowding, past satisfaction) derived from the agent’s belief module, and $\gamma > 0$ controls

distance sensitivity. The destination is sampled from the normalized distribution over candidates:

$$P(j \mid i, t) = \frac{w_{i,j}^t}{\sum_{k \in \mathcal{N}_i^t} w_{i,k}^t}, \quad (2)$$

where $\mathcal{N}_i^t$ denotes the candidate POI set.

3.2.2 Transport Selection

After selecting a destination, the agent chooses a transport mode from available options: {walking, bicycle, car, bus, or train}, based on distance, time of day, weather, urgency, cost, accessibility, persona, and financial pressure.

3.3 Social Interaction

Following CitySim (Bougie and Watanabe, 2025a), agents engage in both face-to-face and online interactions. Face-to-face interactions occur when agents are co-located, and the interaction is compatible with their intention, schedule, and needs, while online interactions occur during leisure or when social need is low. Partner selection is based on relationship strength and recent interaction history. Interaction outcomes update social beliefs (affinity, trust, and familiarity), which in turn shape future decisions such as whether to meet a contact.

3.4 Experience-Driven Reflection

At the end of each day, this module extracts what the agent has learned from recent experiences and converts it into reusable behavioral insights. Rather than simply summarizing logs, the agent reflects on outcomes by considering what it liked or disliked, what constraints shaped its choices, and which habits, preferences, or tendencies emerged. These insights are retrieved and compared with the agent’s existing reflective memory, then integrated as gradual updates so that learned patterns reflect recurring evidence rather than isolated observations. Each reflection records the context in which a tendency applies, the behavioral tendency itself, supporting evidence, and its implications for future behavior. For example, an agent may learn to prefer quiet restaurants after work or leave early to avoid congestion.

3.5 Population-Level Behavioral Alignment

LLM-based simulators are fundamentally limited by their reliance on prompting. Without explicit grounding, agents may reproduce the LLM’s priors about human behavior rather than the population

they are intended to simulate. We propose a calibration mechanism that steers agent decisions toward population-level distributions through textual adapters, without modifying the per-agent reasoning loop or retraining the underlying LLM.

State. For each agent u , we learn a natural-language adapter $\phi^{u,m}$ for each module $m \in \mathcal{M}$ that produce actions. Let Φ_t denote the collection of all adapters at iteration t . In the initial state Φ_0 , all adapters are empty and the simulator reduces to the vanilla agent prompts.

Objective. Let $\mathcal{K}$ index the set of population-level measures, and let $\delta_k(\Phi_t)$ denote the discrepancy between the simulated and target value of measure k under adapter collection Φ_t . We aggregate normalized discrepancies with a geometric mean, which encourages balanced alignment across measures and prevents a single metric from dominating the objective:

$$R(\Phi_t) = \left(\prod_{k \in \mathcal{K}} \frac{\delta_k(\Phi_t)}{\delta_k(\Phi_0) + \varepsilon} \right)^{1/|\mathcal{K}|} \cdot P_{\text{plaus}} \cdot P_{\text{div}}. \quad (3)$$

To prevent trivial solutions, $R(\Phi_t)$ incorporates two penalties: P_{plaus} penalizes a collapse in individual agent plausibility, while P_{div} penalizes a loss of diversity within the population (see Appendix B.1). Adapters are optimized iteratively, using the step return $r_t = R(\Phi_{t-1}) - R(\Phi_t)$ as the improvement signal at each iteration.

Search. Evaluating an adapter update requires running a simulation, hence exhaustive search is infeasible. We formulate adapter optimization as a tree search problem and use Monte Carlo Tree Search (MCTS) (Browne et al., 2012). At each node expansion, an analyst LLM examines the current gap between simulated and target population statistics and proposes a set of possible actions $(m^*, \mathcal{G}, \Delta, g)$. MCTS maintains value estimates for explored branches, so promising edits discovered in earlier branches naturally inform the search in later iterations (Appendix B.3).

Action and Transition. After MCTS selects an action $a_t = (m^*, \mathcal{G}, \Delta, g)$, the tuple specifies a target module m^* , a group of agents $\mathcal{G}$ defined by persona attributes, a behavioral edit Δ , and a short rationale g . Δ is drawn from a vocabulary of behavioral adjustments (Appendix B.2). For each agent in $\mathcal{G}$, a rewriting LLM updates the corresponding adapter using the agent’s current adapter, persona, Δ , and g . The rewriter is constrained

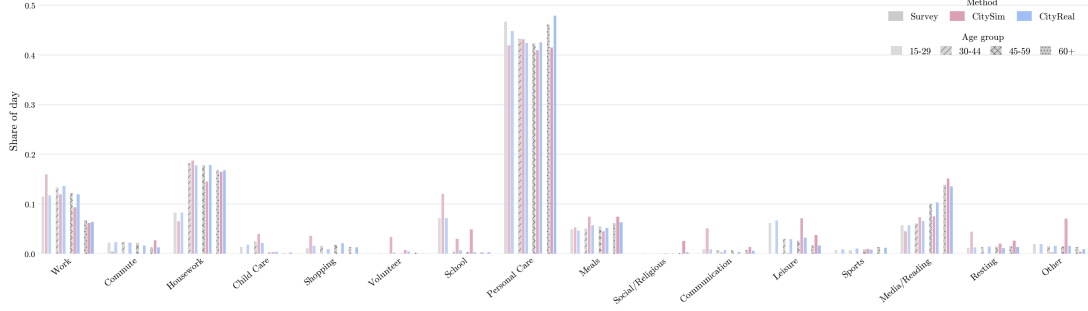


Figure 2: Time-use distribution across activity categories and age groups. Solid bars represent ground truth; striped bars show results from our simulation.

to produce soft behavioral tendencies. Outputs containing clock times, exact distances, or absolutisms (*always*, *never*, *must*) are rejected and regenerated, so adapters bias the agent’s decisions without overriding its situational context and internal states. Adapters for agents outside $\mathcal{G}$ and for modules other than m^* are held fixed, so each action induces a sparse perturbation of Φ_t .

Transfer. To avoid running search over the full population, adapters are optimized on a subset $\mathcal{U}_s$ chosen to cover the main persona types. The learned adapters are then transferred to the *uncalibrated* agents $\mathcal{U} \setminus \mathcal{U}_s$ by nearest-neighbor matching in persona embedding space.

4 Experiments

Settings. All agents are powered by the GPT-5.4-mini version of ChatGPT, except when specified differently, with the number of agents set to 3,000 located in the Tokyo metropolitan area. Unless stated otherwise, adapter search uses only population-level time-use and mobility statistics. Results on these statistics therefore evaluate calibration quality, whereas human-likeness, POI popularity, crowd density, well-being, belief estimation, and counterfactual analyses are excluded from adapter search and serve as evaluations. Appendix A.1 details the corresponding data sources and usage.

Baselines. We compare CityReal with GeAn (Park et al., 2023b), AGA (Yu et al., 2024), HumanoidAgent (Wang et al., 2023b), and MobileCity (Ye et al., 2026). We also report results with our closest competitors, AgentSociety (Piao et al., 2025), and CitySim (Bougie and Watanabe, 2025a).

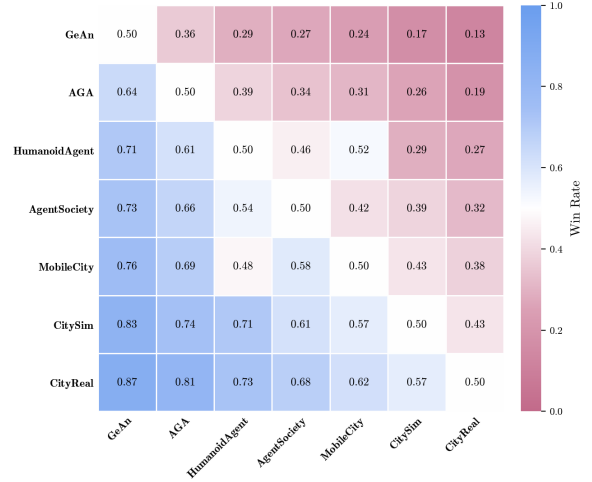


Figure 3: Pairwise win rate matrix. Each entry denotes the proportion of trials in which the row agent is judged more human-like than the column agent.

4.1 Macro-level Time Use

We assess whether CityReal reproduces realistic population-level activity patterns by comparing simulated time use against the 2021 Japanese national time use survey (Statistics Bureau of Japan, 2021). Agents run for two simulated months, and their activities are mapped to survey categories (Work, Commute, Housework, Personal Care & Sleep). We report the share of daily time spent in each category by age group. As illustrated in Figure 2, CityReal closely matches the survey distribution, suggesting that the alignment stage calibrates towards the population-level statistics.

4.2 Pairwise Human Preferences

Beyond macro-level alignment, a realistic simulator should also produce human-like behavior at the individual level. Thus, we compare agents through pairwise judgments of anonymized daily routines, using 15 independent trials per approach. Outputs

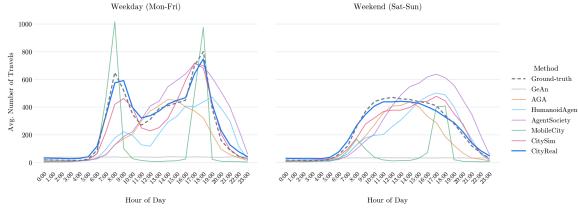


Figure 4: Average number of agent travels per hour on weekdays (left) and weekends (right).

are first normalized with Llama-3.1 70B to mitigate stylistic bias, then judged by GPT-5 along three criteria: (i) **Naturalness**, (ii) **Coherence**, and (iii) **Plausibility**. The pairwise win rate measures how often each agent is judged more human-like than another. Figure 3 shows that CityReal achieves the highest average win rate across pairwise comparisons. This can be attributed to the reflective memory and explicit intention modeling, producing more socioeconomically consistent routines. MobileCity and AgentSociety produce more rigid and repetitive schedules, often overlooking social norms, raising suspicions of AI involvement.

4.3 Travel Patterns

Next, we compare simulated travel distributions against a proprietary city-scale dataset. Figure 4 depicts the average number of trips per hour on weekdays and weekends. Prior methods generate travel decisions mainly from prompt-driven agent decisions. On the other hand, our alignment stage explicitly calibrates agent behavior against observed data. At the individual level, mandatory schedule planning produces consistent commute peaks. As a result, the proposed method closely reproduces both the timing and amplitude of commute peaks on weekdays and the gradual leisure travel patterns on weekends. In contrast, MobileCity produces unrealistically sharp peaks. Other LLM-based approaches capture broad temporal trends but fail to reproduce the precise timing and magnitude of travel peaks.

4.4 Predicting POI Popularity

Predicting which POIs attract the most visitors is central to applications such as urban planning, retail strategy, and event management. We evaluate CityReal as a predictive tool for POI popularity in Shibuya, Tokyo. Ground truth is estimated from Google Maps ratings, while simulated popularity is measured by counting agent visits over one simulated month. We report Spearman rank correlations

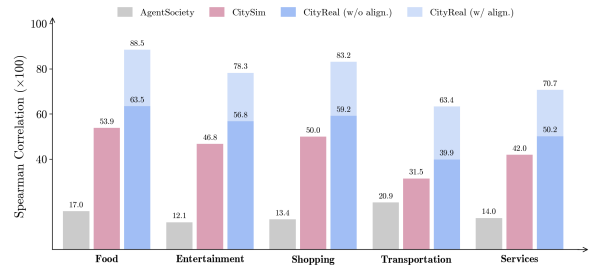


Figure 5: Comparison of real-world POI popularity and simulated visits in Shibuya.

F1-macro (mean $\pm$ std)		
GeAn	0.20 $\pm$ 0.03	
AGA	0.21 $\pm$ 0.04	
HumanoidAgent	0.24 $\pm$ 0.03	
AgentSociety	0.29 $\pm$ 0.02	
MobileCity	0.22 $\pm$ 0.03	
CitySim	0.38 $\pm$ 0.02	
CityReal	0.43 $\pm$ 0.03	
XGBoost	0.45 $\pm$ 0.04	

Table 1: Macro F1-score for well-being class prediction (5-class) across models, evaluated on a proprietary survey. Medals indicate top-3 methods.

between simulated and real-world popularity. Figure 5 indicates that CityReal achieves stronger correlation with real-world popularity than CitySim. One reason is explicit financial constraints, which produce heterogeneous visits across income groups, preventing agents from disproportionately concentrating on highly rated or popular POIs.

4.5 Social Studies using Synthetic Agents

This experiment evaluates CityReal on population well-being estimation using a proprietary dataset of 1,200 survey responses collected in Japan, covering five well-being categories. Agents are initialized with persona profiles matching real respondents and simulate three weeks of daily activity. They then answer the same questionnaire, drawing on their accumulated memories. We compare against an XGBoost baseline trained on real activity and location data, as well as prior agent-based methods. Table 1 highlights that our method outperforms all agent-based baselines, while XGBoost achieves the highest macro F1-score. These findings suggest that agent-based urban simulation can support population-level well-being estimation as a scalable and cost-efficient complement to conventional survey-based analysis.

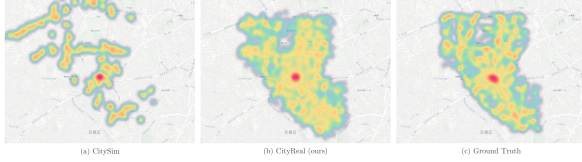


Figure 6: Comparison of simulated (left) and real-world (right) crowd density heatmaps in Shibuya, Tokyo. Warmer colors indicate higher densities.

4.6 Modeling Crowd Density

Predicting spatial crowd density is central to urban management, public safety, and event planning. We assess CityReal’s ability to reproduce pedestrian concentration patterns across Shibuya, Tokyo. Agent visits are aggregated by location to generate simulated density heatmaps, which are compared against ground-truth distributions estimated from smartphone location data. Importantly, the crowd-density signal and Shibuya-specific mobility statistics are not included in the adapter search. As shown in Figure 6, CityReal accurately reproduces real-world density patterns, with the highest concentrations around the train station and major commercial streets. CITYREAL reaches 77.7% spatial overlap with the observed density distribution, compared with 45.3% for CITYSIM. Moreover, 32.8% of simulated visits occur within 300 m of Shibuya station, close to the observed 27.6%, whereas CITYSIM concentrates 60.0% of visits in this area. This is partly driven by the intention module, which keeps agents within a target area and encourages sequential visits to nearby POIs, rather than independent destination choices after each activity. Experience-driven updates further diversify visitation patterns by making future choices depend on past outcomes rather than only on LLM priors.

4.7 Urban Policy A/B Testing

Finally, we examine whether CITYREAL supports counterfactual scenario analysis by simulating two Shibuya (Tokyo) scenarios under sunny and rainy conditions. Because the scenarios differ only in weather, changes in the simulated outcomes reflect how agents adapt their plans to environmental conditions. As highlighted in Figure 7, rain shifts agents away from outdoor discretionary activities toward indoor, covered, or stay-home alternatives, while mandatory routines such as work and commuting remain comparatively stable. This scenario illustrates the practical use of CITYREAL as

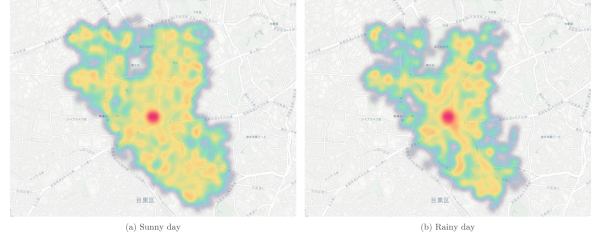


Figure 7: Counterfactual weather analysis in Shibuya. Using the same agents and environment, we compare sunny and rainy conditions. Rain shifts activity from outdoor leisure and walking toward indoor, stay-home, and station-adjacent areas.

a counterfactual planning tool. By estimating how weather changes demand, crowding, and movement patterns, the framework can help planners and operators evaluate event plans, station-area congestion, retail foot traffic, and infrastructure stress under alternative urban conditions.

5 Conclusion

We introduced CityReal, a modular framework for simulating human-aligned urban behavior with LLM-powered agents. Unlike prompt-driven agent simulators, CityReal targets both individual-level coherence and population-level alignment. It maintains intentions across consecutive decisions, allows agents to evolve from experience, and calibrates behavior-generating modules to match real-world statistics through learned adapters. Results demonstrate that our agents closely align with their human counterparts at both micro and macro levels. CityReal enables the study of complex urban phenomena and supports more realistic agent behaviors than prior agent-based models. These findings highlight our approach as a robust foundation for research and industry applications at the intersection of behavioral modeling, urban planning, and social studies.

6 Limitations

Although CityReal achieves strong performance across tasks, several limitations remain. First, the reproducibility of some experiments is constrained by the use of non-public datasets, which may limit direct comparison and independent verification. However, because most evaluations are based on macro-level aggregate statistics, future methods can still compare against CityReal using the reported targets. Second, our framework relies on large language models, hence the generated behav-

iors may reflect cultural, gender, socioeconomic, or other biases present in the underlying models and their training data. Relatedly, we observed occasional hallucinations when agents generated appraisals of recent, uncommon, or less popular POIs, which may introduce errors into downstream simulation outcomes. Third, the quality of the simulated behavior depends on the capabilities and failure modes of the LLM used as the agent backbone. Besides, the alignment process itself may introduce or reinforce biases if the target statistics, calibration data, or textual adapters encode incomplete or skewed assumptions about the population being simulated. Finally, CityReal consists of multiple interacting modules, making it challenging to fully isolate the contribution of each component. We provide ablation studies in the Appendix to partially address this issue, but a more fine-grained analysis of module interactions remains an important direction for future work.

7 Ethics Statement

This paper presents an LLM-driven framework for simulating urban human behavior at scale. Such simulations can support the study of city dynamics, mobility patterns, and social behavior in a scalable and cost-effective way. However, the use of synthetic urban agents also introduces important ethical considerations.

First, synthetic agents may reproduce or amplify biases related to age, gender, occupation, income, lifestyle, or other demographic factors if such biases are present in the underlying LLM, the agent initialization process, or the data used for calibration. If simulation outputs are used to inform urban planning or policy decisions, these biases could lead to analyses that overrepresent some groups while underrepresenting or disadvantaging others. Second, large-scale simulations of human behavior may be used to identify behavioral patterns that could support interventions aimed at steering collective behavior. Without appropriate transparency and oversight, such uses could raise concerns about consent, autonomy, and the potential manipulation of residents or communities.

Third, while synthetic agents are useful for early-stage exploration and low-cost evaluation of urban scenarios, they should not be treated as a substitute for real residents, stakeholders, or domain experts. Simulated behavior necessarily abstracts away many aspects of human experience, including

lived experience, local knowledge, and contextual factors that may not be captured by an LLM-based system. We therefore recommend that synthetic agents be used to complement, rather than replace, participatory design, empirical studies, and expert review, especially when simulation results may influence real-world policies or interventions.

By making these limitations explicit, we aim to encourage the responsible use of LLM-based urban simulations. In particular, we believe such systems should be developed and deployed with attention to transparency, bias evaluation, human oversight, and the social consequences of decisions informed by synthetic populations.

References

- Nicolas Bougie, Gian Maria Marconi, Xiaotong Ye, and Narimasa Watanabe. 2026a. Beyond offline a/b testing: Context-aware agent simulation for recommender system evaluation. *arXiv preprint arXiv:2604.09549*.
- Nicolas Bougie, Gian Maria Marconi, Tony Yip, and Narimasa Watanabe. 2026b. Alignuser: Human-aligned llm agents via world models for recommender system evaluation. *arXiv preprint arXiv:2601.00930*.
- Nicolas Bougie and Narimasa Watanabe. 2024. Generative adversarial reviews: When llms become the critic. *arXiv preprint arXiv:2412.10415*.
- Nicolas Bougie and Narimawa Watanabe. 2025a. Citysim: Modeling urban behaviors and city dynamics with large-scale llm-driven agent simulation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 215–229.
- Nicolas Bougie and Narimawa Watanabe. 2025b. Simuser: Simulating user behavior with large language models for recommender system evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 43–60.
- Nicolas Bougie, Xiaotong Ye, Gian Maria Marconi, and Narimasa Watanabe. 2026c. *Perceptui: Llm agents as human-aligned synthetic users for ui/ux evaluation*. Preprint, arXiv:2606.05697.
- Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Bohnhofshagen, Stephen Tavenor, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43.

- Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.
- Joshua M Epstein. 1999. Agent-based computational models and generative social science. *Complexity*, 4(5):41–60.
- Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. 2025. 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*, pages 5413–5424.
- Atsunori Fujita, Vincent Gong, Masaya Ogushi, Kotaro Yamamoto, Yoshi Suhara, Dane Corneil, and Yev Meyer. 2025. *Nemotron-Personas-Japan: Synthetic personas aligned to real-world distributions*.
- Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1259.
- Jake M Hofman, Duncan J Watts, Susan Athey, Filiz Garip, Thomas L Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J Salganik, Simine Vazire, and 1 others. 2021. Integrating explanation and prediction in computational social science. *Nature*, 595(7866):181–188.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pages 23247–23275.
- David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, and 1 others. 2009. Computational social science. *Science*, 323(5915):721–723.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008.
- Charles M Macal and Michael J North. 2005. Tutorial on agent-based modeling and simulation. In *Proceedings of the Winter Simulation Conference, 2005.*, pages 14–pp. IEEE.
- Saul McLeod. 2007. Maslow’s hierarchy of needs. *Simply psychology*, 1(1-18):3–4.
- OpenStreetMap contributors. 2025. [Openstreetmap](https://www.openstreetmap.org/).
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023a. Generative agents: Interactive simulators of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023b. Generative agents: Interactive simulators of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 52.
- Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, and 1 others. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society.
- Statistics Bureau of Japan. 2021. Survey on time use and leisure activities, 2021. <https://www.e-stat.go.jp/en/stat-search/files?page=1&toukei=00200533&tstat=000001158160>. Accessed July 2024.
- Jiawei Wang, Renhe Jiang, Chuang Yang, Zengqing Wu, Makoto Onizuka, Ryosuke Shibasaki, Noboru Koshizuka, and Chuan Xiao. 2024. Large language models as urban residents: An llm agent framework for personal mobility generation. *Advances in Neural Information Processing Systems*, 37:124547–124574.
- Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023a. Humanoid agents: Platform for simulating human-like generative agents. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 167–176.
- Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023b. Humanoid agents: Platform for simulating human-like generative agents. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 167–176.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Uri Wilensky and William Rand. 2015. *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo*. MIT press.

- Xiaotong Ye, Nicolas Bougie, Toshihiko Yamasaki, and Narimawa Watanabe. 2026. Mobilecity: An efficient framework for large-scale urban behavior simulation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 287–303.
- Yangbin Yu, Qin Zhang, Junyou Li, Qiang Fu, and Deheng Ye. 2024. Affordable generative agents. *arXiv preprint arXiv:2402.02053*.
- Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. 2022. The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science advances*, 8(18):eabk2607.

A Experimental Setup

Figure 1 presents an overview of CityReal. All agent attributes in the persona module are initialized from a proprietary survey-based dataset conducted in Japan. The attribute distributions closely match those observed in recent Japanese census statistics and lifestyle surveys (Statistics Bureau of Japan, 2021). Each persona includes demographic attributes, household composition, life stage, occupation, salary, home and workplace or school locations, and psychographic traits. Big Five personality traits are discretized on a 3-point scale (1=low, 2=medium, 3=high). Home and workplace or school locations are assigned according to Japanese population density and OpenStreetMap data (OpenStreetMap contributors, 2025), ensuring realistic spatial distributions and feasible commutes. For reproducibility, researchers without access to our proprietary persona dataset may alternatively initialize agents from the open NEMOTRON-PERSONAS-JAPAN dataset (Fujita et al., 2025), which provides similar synthetic Japanese personas grounded in real-world demographic, geographic, and personality-trait distributions.

CITYREAL calibrates agent behavior using population-level textual adapters, see Figure 8. Before running the full simulation, the alignment procedure searches for concise adapters attached to behavior modules such as daily planning, intention formation, place selection, transport selection, and reflection. These adapters are soft instructions that steer decisions toward target population statistics without overriding each agent’s persona, memory, needs, location, social context, or financial state. Once learned, adapters are reused across agents, simulated days, and counterfactual scenarios. Note that the orchestrator module is not included in the search.

The temporal memory retrieves the top $k_1 = 5$ entries from the past $\Delta t = 24$ hours using cosine embedding similarity. Each memory entry stores {time, location, activity, observation, outcome, key}. Spatial memory stores beliefs about visited places, including affordability, convenience, crowding, and enjoyment. Beliefs for unvisited POIs are initialized from similar known places, using the $k = 10$ most similar visited locations based on embedding distance. After each visit, the belief module produces a structured feedback of the experience and updates the corresponding spatial belief. Social beliefs over affinity, trust, and familiarity

are updated after interactions.

Agents track short-term needs, including hunger, energy, safety, social connection, and financial security. Need thresholds for action interruption are $T_{\text{hunger}} = 0.3$, $T_{\text{energy}} = 0.3$, $T_{\text{safety}} = 0.2$, and $T_{\text{social}} = 0.2$, with priority order *hunger* > *safety* > *energy* > *social*. Financial security is modeled separately from the immediate interruption needs. Rather than using a fixed interruption threshold, we represent it as a continuous financial pressure signal that conditions activity choice, destination choice, transport selection, and stay-home decisions. Let B_u denote agent u ’s monthly discretionary budget, S_u^t its cumulative discretionary spending up to time t , and $A_u^t = \max(B_u - S_u^t, 0)$ its remaining available budget. Let $\rho_t \in [0, 1]$ denote the fraction of the month elapsed. We define financial pressure as:

$$p_u^t = \text{clip}\left(\frac{S_u^t}{B_u + \epsilon} - \rho_t, 0, 1\right), \quad (4)$$

where higher values indicate that the agent is spending faster than expected for the current point in the month. Similarly, pressure increases when the remaining budget A_u^t is lower than the expected remaining budget $(1 - \rho_t)B_u$. This signal biases agents toward lower-cost POIs, cheaper transport modes, fewer paid leisure activities, or staying at the current location, but it does not directly interrupt ongoing behavior.

The simulation operates with a 5-minute timestep, and all random seeds are fixed for reproducibility. Note that frontier LLM APIs may still exhibit minor nondeterminism due to backend-level implementation details. In detail, daily schedules are constructed from time blocks with a minimum granularity of 5 minutes, matching the resolution of human routine reporting in time-use surveys.

For pairwise human preference evaluation, we define the following criteria: (i) *Naturalness*: the extent to which actions align with the agent’s profile, habits, constraints, and context; (ii) *Coherence*: the logical progression and goal-directedness of activities across time; and (iii) *Plausibility*: the overall believability of the sequence given realistic urban behavior.

A.1 Data Roles in Alignment and Evaluation

Table 2 distinguishes data used for agent initialization, statistics included in the adapter-search objective, and signals used only for evaluation. Results on calibration targets therefore measure whether

adapter search achieves its intended population-level alignment objective, rather than held-out generalization. The underlying LLM remains frozen throughout, with alignment modifying only the textual behavioral adapters. Human-likeness, POI popularity, crowd density, well-being, belief estimation, trip purpose, and counterfactual evaluations use signals that are excluded from adapter search.

All main experiments use geographically consistent Japanese data sources to avoid mixing statistics with different geographic or demographic definitions. For transfer experiments, we instead use local public statistics from the target population (Appendix E.9). When local calibration data are unavailable, the corresponding modules can remain unaligned or reuse transferred adapters, although the cross-city results show that local recalibration performs substantially better.

A.2 Module Details

We provide additional implementation details for the main modules of CITYREAL.

A.2.1 Planning and Intention Module

Daily planning follows a recursive block-based procedure. Starting from an empty day, the planner first assigns fixed activities, such as sleep, work, school, medical appointments, or scheduled commitments, using the agent’s persona, occupation, needs, and constraints. If a selected activity does not fill the entire interval, the block is subdivided according to the activity duration. The planner then recursively fills remaining [EMPTY] blocks with routine activities, such as meals, hygiene, and essential errands. Similar to human routines, flexible blocks are not fixed in advance. When a flexible block begins, the agent forms an intention: a compact natural-language description of what it aims to accomplish, optionally including an area anchor, expected duration, and completion status. Examples include *run errands near the station* or *find an inexpensive meal before returning home*. The active intention provides shared context for subsequent decisions and prevents flexible activities from being generated as disconnected one-step choices.

During simulation, the orchestrator is invoked whenever a new decision is required, such as at the start or end of an activity, upon arrival at a destination, when a need becomes urgent, or when the environment changes. It receives the agent’s current schedule block, active intention, location, retrieved

memories, needs, financial pressure, nearby POIs, nearby agents, and weather. The orchestrator then decides whether to continue, revise, complete, or replace the current intention. The resulting action and observation are stored in temporal memory and used to update spatial beliefs, social beliefs, needs, and financial state.

For place selection, we use the two-stage area-POI procedure described in Section 3.2.1. In all experiments, the area selector considers the top 10 candidate areas by relevance and proximity, and the POI selector considers up to 200 candidate POIs within the selected area. The belief-aware gravity model uses $\gamma = 2.0$ for distance decay. The feature vector $f_{i,j}^t$ contains normalized estimates of affordability, crowding, satisfaction, and convenience from the agent’s spatial memory. For unvisited POIs, these features are averaged over similar visited locations.

For transport selection, available modes encompass: walking, bicycle, car, bus, and train. The transport module conditions on distance, expected travel time, weather, urgency, cost, accessibility, persona constraints, and financial pressure. Face-to-face interactions are limited to one partner per 30-minute window to avoid unrealistically frequent social behavior. Online interactions may occur during leisure periods or when agents seek social contact without traveling.

A.2.2 Reflection Module

At the end of each simulated day, CITYREAL applies the reflection module once for each agent. The input consists of a compact summary of the day’s trajectory, up to 24 representative temporal memories from that day, and the top-8 most relevant existing reflective memories retrieved by semantic similarity. The model generates up to four candidate reflections using a fixed schema: the context in which a tendency applies, the behavioral tendency itself, supporting evidence, and its implication for future behavior. Each candidate is compared with the retrieved reflective memories. If it is consistent with an existing reflection, the entries are consolidated by refining the applicable context and adding the new evidence. If it captures a distinct recurring pattern, it is stored as a separate reflection. At most three new reflective entries are added per day; candidates with limited evidence are not added as standalone entries. During simulation, downstream modules retrieve the top-5 relevant reflective memories. To keep memory bounded, each

Statistic / attribute	Role	Search
Persona attributes	Initialization	–
Home / workplace / school anchors	Initialization	–
Time-use distribution by activity category	Calibration + evaluation	✓
Travel-distance distribution	Calibration + evaluation	✓
Inter-trip interval	Calibration + evaluation	✓
Trip-start (hourly trips)	Calibration + evaluation	✓
Transport mode	Calibration + evaluation	✓
Wake-time	Calibration + evaluation	✓
Trip purpose	Evaluation	✗
POI popularity	Evaluation	✗
Crowd density	Evaluation	✗
Well-being	Evaluation	✗
Human-likeness	Evaluation	✗
Belief estimation	Evaluation	✗
Weather A/B analysis	Scenario analysis	✗

Table 2: Roles of the statistics and attributes used in CITYREAL. “Search” indicates whether the corresponding statistic is included in the adapter-search objective.

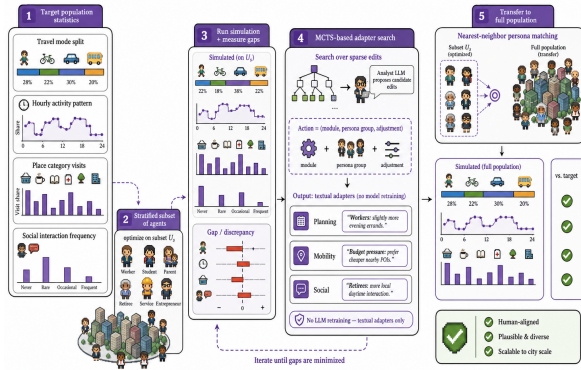


Figure 8: Population-level alignment via MCTS adapter search. An analyst LLM proposes edits to targeted persona groups; a rewriter proposes new textual adapters. A rollout on the subset of agent scores the new adapters, and MCTS uses this score to pick the next edit.

agent maintains at most 100 reflective entries, with older or redundant entries summarized when the limit is exceeded.

B Population-Level Alignment

CITYREAL aligns agent behavior with population-level statistics through textual adapters (Figure 8). For each agent u and behavior module m , the adapter $\phi^{u,m}$ is a short natural-language instruction appended to the module prompt. Let Φ_t denote the collection of adapters at search iteration t . In the initial state Φ_0 , all adapters are empty and the simulator reduces to the unadapted version. The alignment objective compares simulated aggregate statistics with target population statistics.

Let $\mathcal{K}$ denote the set of available measures and let $\delta_k(\Phi_t)$ be the discrepancy between the simulated and target value for measure k . We optimize the baseline-normalized objective:

$$R(\Phi_t) = \left(\prod_{k \in \mathcal{K}} \frac{\delta_k(\Phi_t)}{\delta_k(\Phi_0) + \varepsilon} \right)^{1/|\mathcal{K}|} \cdot P_{\text{plaus}} \cdot P_{\text{div}}, \quad (5)$$

where P_{plaus} penalizes loss of individual plausibility and P_{div} penalizes collapse in population diversity. Lower values of $R(\Phi_t)$ indicate better alignment. The step return is defined as $r_t = R(\Phi_{t-1}) - R(\Phi_t)$, so positive return corresponds to a reduction in discrepancy.

B.1 Alignment Objective Details

Per-Measure Discrepancy. The discrepancy term δ_k in Eq. (3) depends on the type of population-level measure. For continuous distributions, such as travel distance and inter-trip interval, we use a log-space 1-Wasserstein distance combined with a weighted CCDF L_1 term. For categorical distributions, such as transport distribution, origin–destination flows, and trip purpose, we use Jensen–Shannon divergence. For scalar summaries with dispersion, such as wake-up time, we average the relative error of the median and interquartile range.

Plausibility Penalty. This penalty prevents adapters from improving aggregate alignment by making individual agents behave unrealistically. Agents in the simulator emit an *emergency-break*

event whenever a critical need (sleep, food, health) forces them to abort their current activity. Let ρ_t be the mean number of these events per agent-day under adapter collection Φ_t , and ρ_0 the corresponding value under the unadapted simulator Φ_0 . We set the threshold relative to the baseline, $\rho_{\text{thr}} = \max(1.5, \rho_0 + 1.0)$, and apply the penalty:

$$P_{\text{plaus}} = 1 + \lambda_p \cdot \max\left(0, \frac{\rho_t - \rho_{\text{thr}}}{\rho_{\text{thr}} + \varepsilon}\right) \quad (6)$$

with $\lambda_p = 0.5$. Since lower $R(\Phi_t)$ is better, this term raises the objective whenever alignment causes excessive interruptions.

Diversity Penalty. The diversity penalty prevents adapters from matching aggregate statistics by making agents behave too similarly. We partition the population into *groups*. Groups are constructed from bucketed persona and behavior features, including *age_bin*, *household*, *occupation*, *trips_per_day*, and *local_trip_rate*. To ensure reliable group-level estimates, each group must contain at least five agents. For each group and module, we compute the Shannon entropy of agent decisions under the adapted simulator Φ_t and divide it by the corresponding entropy under the unadapted simulator Φ_0 . We then take the geometric mean of these entropy ratios across groups and modules, denoted by h_t . When $h_t < 0.80$, we apply the penalty $P_{\text{div}} = 1 + \lambda_d \max(0, 0.80 - h_t)$, with $\lambda_d = 1.0$. Since $R(\Phi_t)$ is minimized, this term increases the objective when adaptation reduces behavioral diversity by more than 20% relative to the unadapted agents.

B.2 Adapter Edit Vocabulary

Table 3 lists examples of the edit axes used by the analyst during adapter search. Each search action selects one target module and up to two axis-direction pairs. The rewriter then converts the selected edit into a textual adapter.

B.3 Search Procedure

Algorithm. At each iteration, we select a node from the root using UCT, expand it, evaluate each candidate through a rollout, and back-propagate cumulative returns along the evaluated trajectory (Figure 9). All evaluated candidates are added as sibling nodes, so candidates that are not selected immediately can still be revisited by UCT in later iterations. We maintain global statistics for each (module, edit) pair, allowing effective edits to generalize across agent groups.

Module	Axis	Directions
daily structure	regularity	lower / higher
	wake_shift	earlier / later
	evening_activity_tendency	lower / higher
	weekday_rigidity	lower / higher
intention	outing_propensity	lower / higher
	locality_preference	lower / higher
	routine_preference	lower / higher
	social_outing_tendency	lower / higher
trip	exploration_tendency	lower / higher
	multi_stop_tendency	lower / higher
	familiar_place_preference	lower / higher
	destination_range_tendency	shorter / longer
vehicle	walking_preference	lower / higher
	transit_preference	lower / higher
	effort_aversion	lower / higher
	weather_sensitivity	lower / higher
home activity	rest_tendency	lower / higher
	chore_tendency	lower / higher
	hobby_tendency	lower / higher
	family_engagement	lower / higher

Table 3: Behavioral edit axes used for adapter edits. Each axis includes a *neutral* option, indicating that the corresponding behavior should remain unchanged

Candidate Generation. At each expansion step, an analyst LLM compares the simulated and target statistics and proposes candidate adapter edits. Each candidate specifies the module and agent group to adjust. We retain only candidates that use the allowed edit axes in Table 3 and cover a sufficiently large group.

Hyperparameters. We utilize UCT constant $\sqrt{2}$, global exploration constant 1.0, candidate pool size 3, maximum tree depth 4, and at most 20 iterations. Search stops early when the best rollout reward fails to improve by more than 0.005 over 5 consecutive iterations. Analyst and rewriter LLMs use

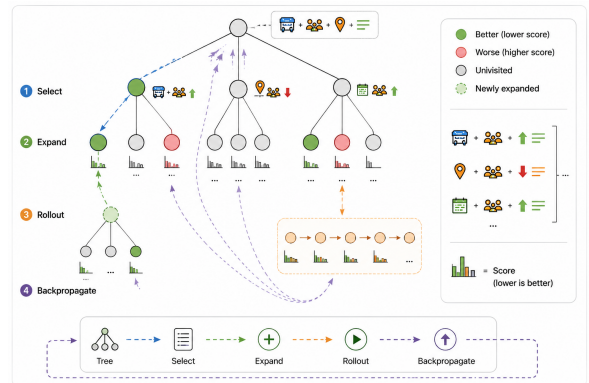


Figure 9: MCTS search procedure.

temperature 0.1 and rollout simulations employ temperature 0.2.

Subset Search. To keep alignment efficient, we optimize adapters on a subset of 500 agents covering major age, household, occupation, and spatial-anchor types. After search, adapters are transferred to the full population by matching each remaining agent to the closest subset agent in persona-embedding space.

C Discussion

We acknowledge that our method exhibits certain limitations. CityReal agents produce collective behaviors consistent with established urban mobility patterns, yet the underlying reasoning of individual decisions remains partially opaque due to the black-box nature of large language models. A plausible explanation is that LLMs internalize behavioral patterns from training corpora covering urban routines and daily life across diverse global contexts. Disentangling learned cultural norms from genuine generalization remains an open challenge for LLM-based simulation.

CITYREAL models socioeconomic constraints through salary-conditioned personas, tracks spending, and estimates a financial-pressure signal. This improves behavioral grounding, but also makes the simulation dependent on the quality of the underlying economic data. When income or expenditure information is coarse, financial pressure may only approximate the constraints faced by real residents, which can bias conclusions about economically driven differences in behavior.

Like other LLM-based simulators, CITYREAL may inherit demographic, cultural, or behavioral biases from the underlying model. Population-level alignment reduces this risk by grounding simulated behavior in observed aggregate statistics, and the financial-state module adds an explicit source of socioeconomic variation. However, aggregate statistics may still hide differences among minority or marginalized groups, especially when these groups are weakly represented in the calibration data. Although our experiments span multiple age groups, occupations, and income levels, evaluating alignment for underrepresented populations remains an important direction for future work.

CityReal uses a hybrid design in which LLMs handle decisions that require contextual reasoning, while explicit models track latent states such as financial pressure, beliefs, and short-term needs.

This design improves both efficiency and interpretability. In preliminary experiments, LLMs were costly and unreliable for estimating scalar internal states, whereas explicit models were easier to inspect and control. Needs are modeled following established psychological frameworks (McLeod, 2007), and financial pressure is computed from tracked income and expenditure, making agent behavior more transparent and auditable.

Some experiments leverage GPT as an evaluator while the agents are generated with models from the same family, which may introduce stylistic bias in pairwise judgments. We mitigate this risk through cross-model validation with Gemini-3.5, but LLM-as-judge evaluation remains imperfect. Future work should incorporate larger-scale human evaluation and additional independent evaluators.

The alignment stage is essential for matching population-level statistics, but it depends on the quality of the calibration targets. If the target statistics are incomplete, biased, or too coarsely aggregated, the learned adapters may reproduce these limitations rather than recover the true behavioral distribution. Optimizing aggregate fit can also downweight rare but realistic behaviors, such as atypical routines or uncommon activity sequences. Although the plausibility and diversity penalties are designed to reduce this risk, preserving long-tail behavior remains challenging. Finally, adapters are tied to the target population and city context, and may need to be recalibrated when transferred to new urban environments. Cross-city experiments in Appendix E.9 show that the underlying architecture transfers to Paris and New York, but also confirm that target-city recalibration is substantially more effective than transferring Tokyo adapters unchanged.

CITYREAL also depends on sufficient local data for initialization. When demographic, spatial, or economic records are sparse, personas, spatial anchors, and financial states become less reliable. This weakens downstream modules such as belief formation and reflection, which rely on meaningful prior context and accumulated experience. Salary-informed initialization and empirical demographic distributions improve grounding, but they remain approximations of real individual profiles, especially for groups with limited observational coverage.

Method	Activity	Dialogue	Mobility	Event Reaction
GeAn	3.16 $\pm$ 0.19	3.97 $\pm$ 0.04	3.13 $\pm$ 0.16	3.06 $\pm$ 0.21
AGA	3.24 $\pm$ 0.27	3.95 $\pm$ 0.03	3.26 $\pm$ 0.24	3.17 $\pm$ 0.20
HumanoidAgent	3.38 $\pm$ 0.31	4.00 $\pm$ 0.06	3.30 $\pm$ 0.21	3.31 $\pm$ 0.17
AgentSociety	4.05 $\pm$ 0.23	4.09 $\pm$ 0.05	3.88 $\pm$ 0.25	3.86 $\pm$ 0.21
MobileCity	4.14 $\pm$ 0.26	4.05 $\pm$ 0.06	4.06 $\pm$ 0.19	3.83 $\pm$ 0.17
CitySim	4.40 $\pm$ 0.17	4.25 $\pm$ 0.04	4.23 $\pm$ 0.15	4.15 $\pm$ 0.15
CityReal w/o Align.	4.47 $\pm$ 0.16	4.28 $\pm$ 0.05	4.33 $\pm$ 0.15	4.27 $\pm$ 0.14
CityReal	4.56 $\pm$ 0.14	4.32 $\pm$ 0.05	4.46 $\pm$ 0.13	4.41 $\pm$ 0.13

Table 4: Human-likeness scores evaluated by GPT-5 across city simulation domains. Higher values indicate greater similarity to real human responses.

D Pseudo-Code

We summarize the main simulation loop of CITYREAL below:

Algorithm 1: CITYREAL Simulation Loop

```

Input: city environment, agents, target population statistics
Output: trajectories and aggregate urban outcomes
calibrate_modules() // population-level textual adapters
initialize_agents() // personas, needs, budgets, beliefs, memories
For each day:
  For each agent:
    plan_day() // mandatory, routine, flexible blocks
  For each time step:
    For each active agent:
      perceive(); retrieve_memory()
      update_needs_and_financial_pressure()
      decision ← orchestrator()
      // continue, revise, complete, or replace intention
      if decision.requires_new_activity:
        intention ← update_intention()
        activity ← select_activity(intention)
        if activity.requires_move:
          area, poi ← select_place(intention)
          transport ← select_transport(poi)
          move(poi, transport)
      else:
        continue_current_behavior()
        social_interaction()
        execute_activity()
        update_state() // memory, beliefs, needs, spending
    For each agent:
      reflect_end_of_day() // habits, preferences, constraints, beliefs
      aggregate_outcomes() // mobility, crowds, popularity, well-being

```

E Additional Experiments

E.1 Human Likelihood

Following Chiang and Lee (2023), GPT-5 assesses whether agent behaviors appear human-generated or LLM-generated. For each method, we collect

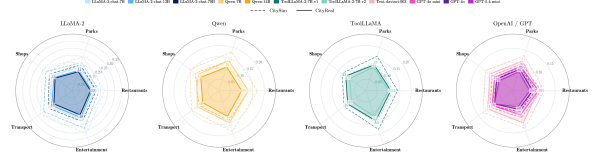


Figure 10: Category-wise mean absolute error (MAE) of belief estimation for unvisited POIs, evaluated across five semantic categories (*Restaurants, Parks, Shops, Transport, Entertainment*) and eleven LLM models. Lower values indicate higher accuracy.

20,000 outputs across four domains: daily activities, dialogue, mobility choices, and event reactions. Each sample is scored on a 5-point Likert scale, with higher scores indicating greater resemblance to human behavior. Table 4 reports that CITYREAL achieves the highest scores across all domains. Gains are largest for mobility and event reaction, where persistent intentions and belief-aware mobility produce more coherent responses to changing contexts. This indicates that population-level adapters provide additional behavioral calibration, including at the agent level. Notably, the variant without alignment also outperforms CITYSIM. Dialogue gains are smaller, which is expected because dialogue fluency is already strong across recent LLM-based simulators. Overall, these results suggest that CITYREAL yields more human-like behavior by making agent decisions more coherent, context-sensitive, and faithful.

E.2 Belief Estimation

Place beliefs directly condition destination choice and activity planning in CityReal. We evaluate belief estimation accuracy by initializing each agent with belief vectors derived from visited POIs, then tasking it with predicting beliefs for a disjoint set of unvisited POIs. For each test POI, we compute the mean absolute error (MAE) between predicted and ground-truth beliefs across five semantic categories. As shown in Figure 10, larger models achieve lower errors overall, with GPT-5.4-mini performing best across categories, followed by GPT-4o mini and Qwen-14B. ToolLLaMA is competitive for *Transport* and *Shops*, while smaller models such as LLaMA-2 produce larger errors, especially for *Entertainment*. This study indicates that stronger models provide more reliable place-belief estimates, which is important because CITYREAL uses these beliefs to condition destination choice, financial decisions, and long-term place preferences.

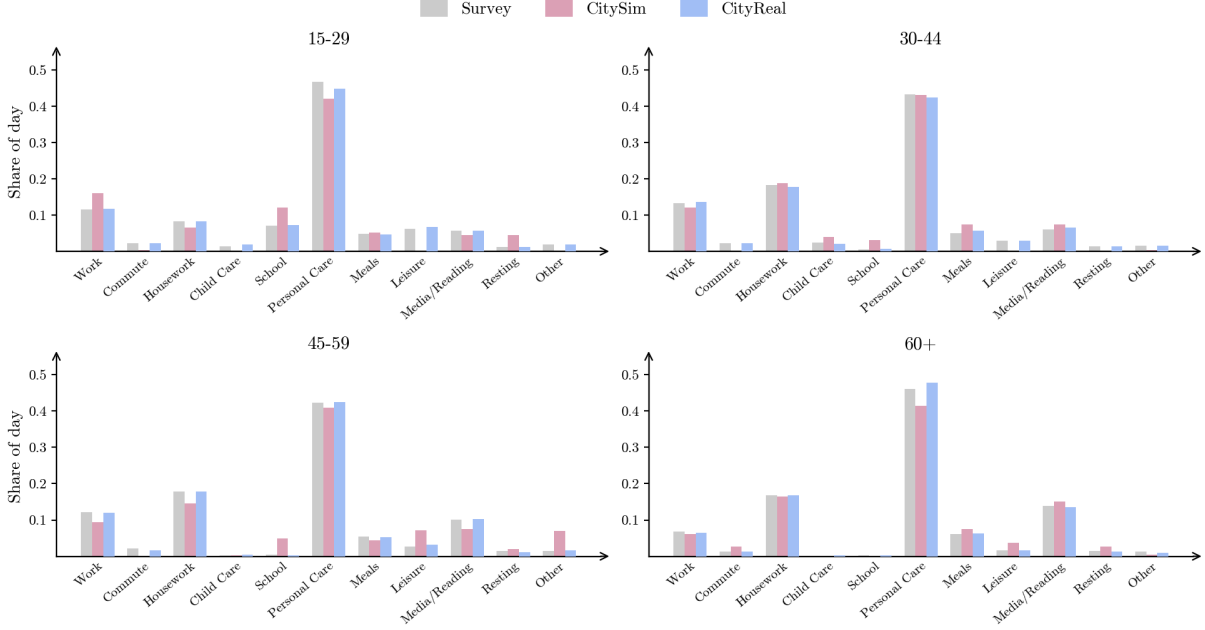


Figure 11: Time-use distributions by age group.

E.3 Population-level Time-use Alignment

This experiment examines whether CITYREAL improves alignment within demographic groups, rather than only matching aggregate population statistics. Agents are grouped by age and compare their simulated time-use distributions with the 2021 Japanese national time-use survey (Statistics Bureau of Japan, 2021). For each bin, we compute the Jensen–Shannon divergence between the simulated and survey activity distributions, where lower values indicate closer alignment. As shown in Figure 11, CITYREAL reduces divergence across all age groups compared with CITYSIM. The alignment stage preserves age-specific behavioral patterns, rather than improving the population average by collapsing heterogeneous groups into a single distribution.

E.4 Adapter Convergence and Efficiency

Next, we evaluate whether the adapter search procedure effectively reduces the population-level alignment objective. Starting from empty adapters, we compare MCTS with two simpler alternatives under the same evaluation budget: random search, which samples adapter edits uniformly, and greedy search, which applies the best local edit at each iteration. We report the best-so-far value of $R(\Phi_t)$ across search iterations, where lower values indicate better alignment. MCTS reaches the lowest final objective, as illustrated in Figure 12. Greedy

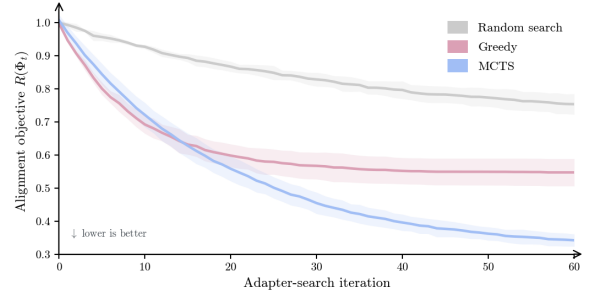


Figure 12: Adapter convergence under the same evaluation budget. We plot the best-so-far alignment objective $R(\Phi_t)$. Lower is better.

search improves quickly at first but then plateaus, suggesting that local edits can correct salient discrepancies but are less effective at coordinating multiple modules and population subgroups. Random search improves more slowly under the same budget. These results support the use of tree search for adapter calibration, as MCTS can explore alternative edit sequences while reusing value estimates from earlier branches.

E.5 Robustness to Evaluator Bias

Our main pairwise evaluation uses GPT to judge which simulated behavior appears more human-like. Since CITYREAL also relies on OpenAI models for agent generation, this setup may raise concerns that the evaluator shares model-family biases with the system being evaluated. To assess this

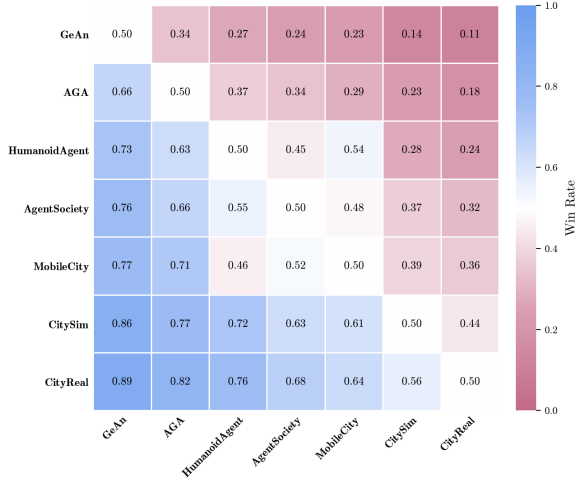


Figure 13: Pairwise win-rate matrix using Gemini-3.5 as the evaluator. Each entry indicates the fraction of comparisons in which the row framework is judged more human-like than the column framework.

risk, we repeat the same win-rate experiment with Gemini-3.5, an evaluator from a different model family. Figure 13 depicts that the overall ranking remains stable: CITYREAL is still preferred over prior frameworks, including CITYSIM. The margin over CITYSIM is modest, but consistent, suggesting that the main human-likeness results are not solely driven by evaluator bias from using an OpenAI judge.

E.6 Ablation Study

This section examines the contribution of each component of CITYREAL. Starting from the full model, we remove one component at a time: population-level alignment, financial pressure, social interaction, experience-driven reflection, intention formation, recursive planning, beliefs, needs, and persona. We evaluate both *individual* realism, measured by human-likeness across four domains, and *population-level* alignment, measured by the Jensen–Shannon divergence between simulated and survey distributions for time use and transport-mode share. Table 5 reports the results. The two evaluation axes capture different aspects of realism. Removing alignment mainly affects population-level metrics, substantially increasing the divergence for both time use and transport-mode share. It also reduces human likeness across all domains, suggesting that the learned adapters improve not only aggregate calibration but also individual behavioral realism by preventing agents from reverting to generic LLM priors. Other ablations keep

the alignment stage active and therefore remain close to the calibrated population targets, while primarily affecting individual coherence. Among the CityReal components, removing financial pressure degrades mobility and transport-mode alignment, as agents lose an important socioeconomic signal for destination and mode choice. Removing intention mainly affects mobility and event reaction, as without persistent motives, flexible activities become disconnected one-step decisions, weakening trajectory continuity. Disabling reflection produces large drops in activity and event reaction, since agents can no longer accumulate habits and preferences from experience. Removing social interaction mainly affects dialogue and event reaction, while activity is less affected. Besides, our alignment technique also reduces human likeness across all domains, suggesting that the learned adapters improve not only aggregate calibration but also individual behavioral realism by preventing agents from reverting to generic LLM priors. In contrast, all other ablations retain the alignment stage and therefore remain relatively close to the target population distributions.

E.7 Effect of Search Subset Size

The alignment search is performed on a subset of agents, and the learned adapters are transferred to the remaining population. Increasing the subset size consistently improves population-level alignment, as reported in Table 6. Small subsets underrepresent minority persona groups, leading to weaker calibration, especially for transport-mode distribution. Gains diminish once the main persona types are sufficiently covered, suggesting that full-population search provides only limited additional benefit. We therefore use 500 agents in our experiments, which achieves most of the alignment improvement while limiting search cost.

E.8 Robustness across LLM Backbones

We evaluate whether the benefits of CITYREAL depend on a specific LLM backbone by repeating the same simulation and alignment procedure with GPT-5.4-mini, Gemini-3.5-Flash, and Claude Haiku 4.5. Individual-level human-likeness is evaluated using the same GPT-5 protocol as Table 4, while population-level alignment is measured by time-use and transport-mode JSD. Table 7 depicts that the framework remains effective across all three backbones. Stronger backbones obtain higher absolute human-likeness, but each model reaches

	Human-likeness (GPT-5, Likert 1–5, ↑)								Population JSD (↓)	
	Activity		Dialogue		Mobility		Event Reaction		Time-use	Transport
CityReal (full)	4.56 ± 0.14	 	4.32 ± 0.05	 	4.46 ± 0.13	 	4.41 ± 0.13	 	0.0016	0.0021
w/o Alignment	4.43 ± 0.17	 	4.27 ± 0.06	 	4.27 ± 0.16	 	4.19 ± 0.15	 	0.066	0.058
w/o Intention	4.08 ± 0.22	 	4.12 ± 0.07	 	3.95 ± 0.23	 	4.00 ± 0.19	 	0.0058	0.0055
w/o Reflection	3.95 ± 0.25	 	4.02 ± 0.07	 	3.97 ± 0.21	 	3.88 ± 0.22	 	0.0044	0.0041
w/o Financial	4.30 ± 0.21	 	4.23 ± 0.06	 	4.13 ± 0.18	 	4.19 ± 0.17	 	0.0049	0.0112
w/o Persona	3.66 ± 0.28	 	3.55 ± 0.12	 	3.72 ± 0.25	 	3.60 ± 0.24	 	0.0091	0.0083
w/o Needs	3.92 ± 0.27	 	4.08 ± 0.08	 	4.02 ± 0.19	 	3.85 ± 0.25	 	0.0063	0.0049
w/o Belief	4.05 ± 0.22	 	4.10 ± 0.09	 	3.98 ± 0.23	 	3.92 ± 0.22	 	0.0038	0.0035
w/o Rec. Plan	4.02 ± 0.24	 	4.14 ± 0.07	 	4.06 ± 0.20	 	4.05 ± 0.21	 	0.0071	0.0038
w/o Social	4.26 ± 0.19	 	3.98 ± 0.08	 	4.10 ± 0.20	 	4.03 ± 0.18	 	<u>0.0026</u>	<u>0.0029</u>

Table 5: Ablation study for CITYREAL. Human-likeness is rated by GPT-5 (Likert, 1–5, mean ± std; higher is better) for activity, dialogue, mobility, and event reaction; medals denote the top-3 configurations per domain. The last two columns report population-level alignment as the Jensen–Shannon divergence between simulated and survey distributions for time use (calendar) and transport distribution (vehicle), where lower is better.

Subset	Cov.	Time-use JSD	Transport JSD
100	3%	0.0072 ± 0.0019	0.0103 ± 0.0026
250	8%	0.0034 ± 0.0011	0.0050 ± 0.0015
500	17%	0.0016 ± 0.0004	0.0021 ± 0.0006
1000	33%	0.0012 ± 0.0003	0.0015 ± 0.0004
3000 (full)	100%	0.0010 ± 0.0002	0.0012 ± 0.0003

Table 6: Effect of alignment search subset size on population-level alignment for a 3,000-agent population. “Cov.” denotes the subset coverage. We report Jensen–Shannon divergence for time use and transport-mode share. Lower is better.

low population-level divergence after alignment, indicating that the architecture and calibration procedure are not specific to a specific backbone.

E.9 Cross-City Transfer

To examine whether CITYREAL transfers beyond Tokyo, we repeat the POI-popularity experiment in Paris and New York. The agent architecture is unchanged across cities. We compare four settings: CITYSIM, CITYREAL without population alignment, CITYREAL with the Tokyo adapters transferred unchanged, and CITYREAL after recalibration using local public population statistics from INSEE for Paris and U.S. public statistics for New York. Table 8 highlights that the architecture itself transfers across cities: even without alignment, CITYREAL outperforms CITYSIM in every POI category. Directly transferring the Tokyo adapters provides only a modest additional improvement and can degrade categories whose behavior differs across cities, particularly transportation. In contrast, recalibration using target-city statistics

substantially improves correlation in both Paris and New York.

E.10 Human Validation of Pairwise Evaluation

Our pairwise evaluation uses GPT-5 to assess the human-likeness of generated trajectories. We validate this protocol with three human annotators using the same criteria and repeat the evaluation with Gemini-3.5. As shown in Table 9, all evaluators preserve the same ranking, with CITYREAL preferred over every baseline. Human annotators show substantial agreement (Fleiss’ $\kappa = 0.67$), while agreement between the human majority and GPT-5/Gemini-3.5 reaches Cohen’s $\kappa = 0.65/0.63$, respectively. These results support the robustness of the pairwise evaluation with LLMs as judges.

F Cost Analysis

We report the LLM cost and scalability of CITYREAL and compare it with CitySim, the closest baseline. Table 10 presents the daily token usage and estimated cost per 1,000 agents, using GPT-5.4-mini pricing. Because CITYREAL includes population-level alignment, we also report the one-time cost of adapter calibration. In our default setting, calibration is performed on a subset of agents with short simulation rollouts, costing approximately \$248. This cost is incurred once for a target population and can be reused across agents, simulation days, and scenario analyses. The recurring daily cost of CITYREAL remains comparable to scalable LLM-agent baselines and substantially lower than prompt-heavy simulators such as GeAn. Costs scale approximately linearly with the agent

Backbone	Activity ↑	Dialogue ↑	Mobility ↑	Event Reaction ↑	Time-use JSD ↓	Transport JSD ↓
GPT-5.4-mini	4.56	4.32	4.46	4.41	0.0016	0.0021
Gemini-3.5-Flash	4.46	4.26	4.38	4.29	0.0024	0.0027
Claude Haiku 4.5	4.38	4.30	4.19	4.18	0.0043	0.0056

Table 7: Robustness across LLM backbones. Human-likeness is evaluated on a 1–5 scale, while the last two columns measure population-level alignment.

Method	Food	Entertainment	Shopping	Transportation	Services	Mean
Paris						
CitySim	47.6	43.4	44.9	29.8	36.5	40.4
CityReal w/o Alignment	55.8	52.4	54.6	34.6	45.3	48.5
CityReal + Tokyo Adapters	61.7	55.1	57.9	33.2	45.6	50.7
CityReal + Paris Recalibration	78.6	68.4	74.5	51.4	63.1	67.5
New York						
CitySim	45.9	44.6	43.0	23.8	36.2	38.7
CityReal w/o Alignment	54.7	54.3	51.8	28.9	43.4	46.6
CityReal + Tokyo Adapters	57.5	53.7	54.1	24.6	45.2	47.0
CityReal + NY Recalibration	76.1	70.9	71.0	44.1	59.4	64.3

Table 8: Cross-city POI popularity prediction in Paris and New York. We report Spearman rank correlation between simulated and real-world POI popularity.

Comparison	Human	GPT-5	Gemini-3.5	Method	Tokens/day (M)	Cost/day (USD)
CityReal vs. GeAn	0.92	0.87	0.89	CitySim	16.71	\$5.51
CityReal vs. AGA	0.84	0.81	0.82	AgentSociety	14.98	\$4.82
CityReal vs. HumanoidAgent	0.74	0.73	0.76	MobileCity	14.41	\$3.46
CityReal vs. AgentSociety	0.72	0.68	0.68	AGA	22.05	\$7.01
CityReal vs. MobileCity	0.63	0.62	0.64	GeAn	69.84	\$23.60
CityReal vs. CitySim	0.59	0.57	0.56	CityReal (simulation only)	19.58	\$6.76

Table 9: Pairwise human-likeness evaluation with human annotators and two LLM judges. Values report the fraction of comparisons in which CITYREAL is preferred over the corresponding baseline.

population, and the daily cost for 1,000 agents remains much lower than involving real humans in comparable social studies.

G Comparison with Prior Work

This section provides a systematic comparison of CITYREAL against recent agent-based urban simulation systems. Table 11 summarizes agent modeling dimensions, including alignment, adaptive memory, intention modeling, financial state, social interaction, and scalability. Table 12 focuses on spatial and mobility reasoning.

Table 10: Daily LLM token usage and estimated cost per 1,000 agents using GPT-5.4-mini.

H LLM Evaluator Prompt

To assess whether interaction traces resemble those of real users or are indicative of AI-generated behavior, we employ an LLM-based evaluator. This judge is prompted as follows:

LLM Evaluator Prompt

Please evaluate the following urban behavior trace of a simulated agent and determine whether it resembles behavior generated by an AI agent or by a real human:

{interaction logs}

Please rate on a scale of 1 to 5, with 1 being most like an AI and 5 being most like a human.

Method	Pop. Align.	Rich Persona	Needs	Economy	Adaptive Memory	Persistent Intent.	Social	Scale
CITYREAL	✓	✓	✓	✓	✓	✓	F2F + online	>10k
CitySim	✗	✓	✓	✗	partial	✗	F2F + remote	>10k
AgentSociety	✗	partial	✓	✗	partial	✗	F2F	>10k
MobileCity	✗	partial	✗	✗	partial	✗	F2F	4k
GeAn	✗	partial	✗	✗	partial	✗	F2F	25
AGA	✗	partial	✗	✗	partial	✗	F2F	100
HumanoidAgent	✗	partial	✗	✗	✗	✗	F2F	100

Table 11: Comparison of agent-level modeling capabilities. CITYREAL is the only framework that jointly supports population-level alignment, persistent intentions, financial state modeling, and adaptive memory at large scale.

Method	Intent.-Aware Area	POI Beliefs	Belief Updates	Belief-Gravity	Transport Choice	Cost-Aware Choice
CITYREAL	✓	✓	✓	✓	✓	✓
CitySim	✗	✓	✓	✓	✓	✗
AgentSociety	✗	✗	✗	✗	✗	✗
MobileCity	✗	✗	✗	✗	rule-based	✗
GeAn	✗	✗	✗	✗	✗	✗
AGA	✗	✗	✗	✗	✗	✗
HumanoidAgent	✗	✗	✗	✗	✗	✗

Table 12: Comparison of spatial and mobility reasoning. CITYREAL extends prior mobility models with intention-aware area selection, belief-aware gravity scoring, transport choice, and cost-aware decision making.

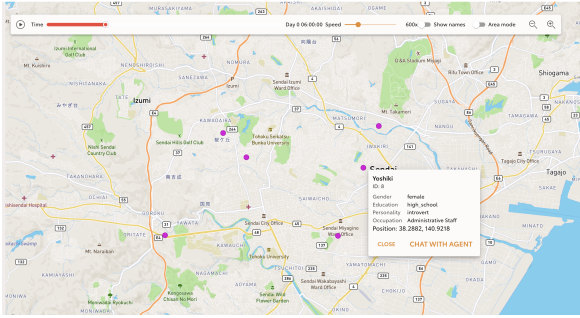


Figure 14: Simulation interface for visualizing agent behavior and city dynamics.

I Simulation Interface

To support qualitative analysis and scenario exploration, we develop an interactive interface for visualizing agent behavior and city-level dynamics. Figure 14 illustrates the interface, which visualizes agent trajectories, activities, intentions, schedules, mobility patterns, and contextual information in real time. It also allows users to inspect individual agents, compare behaviors across populations, and analyze how routines evolve under different environmental conditions or interventions. This interface is designed to support researchers in understanding not only aggregate simulation outcomes, but also the underlying behavioral processes that generate them.

J Reproducibility

For transparency and reproducibility, we release all prompts used in our framework, including those related to modules and alignment procedure in a GitHub repository at https://github.com/simul-ci/city_real. The repository further contains comprehensive documentation covering data preprocessing procedures, hyperparameter configurations, and prompt specifications.