Bridging Individual and Collective Realism in LLM-Based Human Mobility Simulation via Mobility Scaling-Law Guidance

Hua Yan
Lehigh University
Bethlehem, USA
huy222@lehigh.edu

Heng Tan
Lehigh University
Bethlehem, USA
het221@lehigh.edu

Yu Yang
Lehigh University
Bethlehem, USA
yuyang@lehigh.edu

Abstract

Geospatial applications such as urban planning, epidemic forecasting, and transportation demand modeling depend on individual mobility data, but such data are costly to collect, uneven in coverage, and privacy-sensitive. Human mobility simulation offers a scalable alternative. A recent line of work treats large language models (LLMs) as human agents, modeling individual cognitive processes to generate realistic trajectories. Yet because each agent is simulated in isolation, these methods provide no population-level coordination mechanism, and the collective regularities of real mobility — how trip distances, visited locations, and flows distribute across a population — fail to emerge. We close this gap with COMPASS, which turns empirical mobility scaling laws into a feedback signal that guides prompt construction. COMPASS starts from coarse, population-level adjustments driven by these scaling laws and progressively refines them into individual prompts, jointly satisfying multiple aggregate objectives while keeping individual trajectories realistic. Across two public datasets, COMPASS outperforms state-of-the-art LLM-based simulators.

Keywords

Human mobility simulation; Large language model; Scaling law of human mobility

1 Introduction

Human mobility captures how populations move across geographic space over time, providing essential information for geospatial and spatiotemporal applications in location-based services [38, 39], traffic forecasting [17, 18], epidemic forecasting [6, 36], and urban resource allocation [33]. In practice, such data come from travel surveys and sensor-based tracking, but both channels have limitations that cap their scalability. Travel surveys record individual trips in detail, yet their high cost and infrequent deployment limit the population they can cover [3, 31]. Sensor-based tracking, such as mobile-phone traces and Bluetooth beacons, captures finer temporal detail, but its coverage hinges on device adoption and it raises serious privacy concerns [15]. As a result, neither channel readily delivers mobility data that is both population-scale and privacy-preserving.

Recent studies [1, 8, 13, 14, 16, 19, 22, 25] prompt large language models (LLMs) to generate synthetic mobility trajectories in a zero-shot way. Unlike approaches that train a trajectory model on real-world data [9, 21, 44], LLM-based methods enable large-scale data generation while preserving individual privacy. These methods treat the LLM as a human agent: given a user profile, the model reasons step by step about a person's mobility intentions and produces a trajectory accordingly. They focus on making each agent's behavior more human-like by modeling individual-level cognitive processes such as intention and reflection. These methods, however, simulate each agent on its own, with no mechanism to coordinate behavior across the population. Individual trajectories may therefore look plausible while their aggregate fails to reproduce the mobility patterns seen in real data (Section 2.1 provides empirical evidence). This gap matters for downstream tasks such as transportation planning and epidemic modeling, where decisions hinge on accurate population-level flows, behavioral heterogeneity, and rare events rather than on individually plausible routines. Our goal is to *develop an LLM-based mobility simulation framework that keeps individual behavior realistic while also matching the population-level patterns found in real data.*

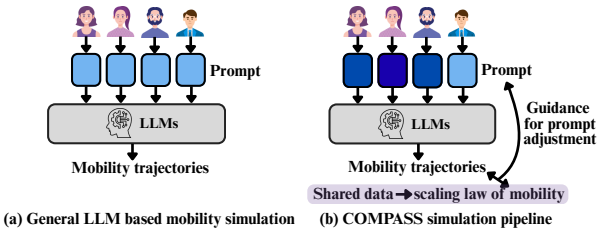

Figure 1: Core idea of COMPASS.

Prior research in social physics shows that human mobility follow stable and reproducible scaling laws at the population level [4, 10, 23, 29]. These laws surface in observable measures such as travel distance and radius of gyration, which remain accessible in public mobility datasets even at coarse spatial or temporal resolution. We argue that these scaling laws of human mobility can serve as population-level guidance for bridging individual-level trajectory generation and collective realism in LLM-based human mobility simulation. For example, the radius of gyration [10] quantifies the spatial range of individual mobility, and has been shown to follow a truncated power law. By comparing the distributions observed in real-world data with those generated by simulation, we can identify differences, such as fewer short-distance trips (around 1 km) and more very long-distance trips (beyond 200 km) in the simulated trajectories. These differences provide diagnostic signals, indicating which types of individuals are not adequately captured by the generative process. Based on this analysis, we can tailor individual-level prompts by incorporating explicit behavioral constraints and persona descriptions, so that different individuals are guided by distinct prompts rather than sharing a single prompt that varies only in profile information. These targeted adjustments correct underrepresented mobility behaviors in the generative process and thereby reproduce the population-level mobility patterns reflected by scaling laws of human mobility.

Building on this idea, we leverage scaling laws of human mobility to guide individual-level prompt adjustment, generating trajectories

that better reproduce population-level patterns, as shown in Figure 1. Such laws are usually estimated from detailed trajectory data, but we find they survive largely intact in coarse-grained shared datasets (details in Section 2.3). This is consistent with emerging data-sharing practices, where what is released is often coarse trajectories [37], validated simulation output [42], or aggregated statistics [10], and it means our approach can operate under limited data access, improving both generalizability and practical deployability.

Using scaling laws this way is not straightforward. Because they manifest across several facets of population behavior, such as spatial range, temporal regularity, and visitation preference, it is unclear which individuals to adjust and which aspects of their behavior to change. Suppose a simulation produces too few short trips and too little activity around midday and at night; this does not mean every adjusted individual should both travel short distances and be active during those hours. Adjustments that satisfy one scaling-law pattern must also be reconciled with the others, since fixing one can easily distort another.

To address these challenges, we view prompt adjustment as a multi-objective, multi-prompt optimization problem, formulated as a Markov Decision Process (MDP) and solved with Monte Carlo Tree Search (MCTS). A state is the current set of prompts for all individuals; an action is an LLM-generated coarse-grained adjustment strategy applied to a group of individuals; and the reward measures how closely the resulting trajectories match the scaling-law objectives. To keep these objectives from competing, we design a multi-objective reward that credits progress on one objective only when it does not degrade the others. Although each strategy is defined at the group level, we design a two-level MCTS procedure to refine its application at both the strategy and user levels: the search selects promising adjustment strategies while identifying the users most responsive to each strategy. Across iterations, individuals accumulate different combinations of adjustments, so coarse group strategies become fine-grained, individual-level adaptations that improve the population-level objectives together. To reduce cost, we introduce a context-aware global action value estimator to prioritize promising adjustment strategies before expensive trajectory regeneration and evaluation. We further perform the search on a small subset of individuals and then generalize the resulting adjustments to larger groups of individuals with similar profiles.

In particular, our main contributions are as follows.

- We explore the idea of leveraging scaling laws of human mobility from shared data as guidance to bridge individual trajectory generation and collective realism in LLM-based mobility simulations.
- We design COMPASS, a COllective Multi-Prompt Adjustment framework guided by Scaling lawS for large-scale LLM-based human mobility simulation. Our framework starts from LLM-generated coarse-grained adjustment strategies and uses a two-level MCTS search to decide both which strategies to apply and which individuals should receive each strategy. By allowing different individuals to undergo different combinations of adjustments, COMPASS turns coarse strategies into fine-grained individual-level prompt adaptations, which jointly improve multiple scaling-law objectives at the population level.
- We conduct extensive experiments on two public datasets, and the results demonstrate that our method achieves the best performance, with improvements ranging from 10.70% to 74.79% over the best baseline across multiple metrics. In addition, we evaluate the impact of different types of shared data on our method. The results show that mobility measures obtained from different types of shared data can effectively guide LLM-based mobility simulations. Notably, even statistical shared data, without trajectory-level information, leads to noticeable performance improvements.

2 Motivation

2.1 Limitations in reproducing scaling laws

We begin with two classical scaling laws of human mobility, those for travel distance and location visitation frequency, to show that existing simulations cannot fully reproduce the population-level behaviors observed in real data. As a representative case, we compare LLMob [13], a recent LLM-based simulator, against real-world data from Beijing (data details in Section 4.1).

Travel distance. Travel distance is the distance Δd between consecutive locations in an individual's trajectory, capturing the spatial scale of human mobility. It follows a truncated power-law, $P(\Delta d) \sim (\Delta d + \Delta d_0)^{-\beta} \exp\left(-\frac{\Delta d}{\kappa}\right)$, where Δd_0 is a small offset parameter [10]. The exponent β determines how frequently long-distance movements occur, with larger values leading to fewer long trips and smaller values leading to more long trips. The cutoff κ captures the spatial extent of individual mobility.

We compute the travel distance from both the real-world data and the simulation and visualize their complementary cumulative distribution functions (CCDFs) in Figure 2. The real-world data are well described by a truncated power-law, with parameter values consistent with those reported in previous studies [10] (with $\beta \approx 1.75 \pm 0.15$ and $\kappa \approx 400$ km). In contrast, for the simulation, the estimated scaling exponent is smaller ($\beta \approx 1.22$), and is accompanied by a cutoff at around ($\kappa \approx 1039.3$ km), which is much larger than that observed in real-world data. These results show that the simulation produces too many long-distance trips and does not capture the shape decline in such trips that the real-world data exhibit.

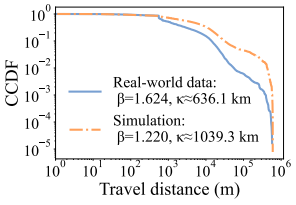

Figure 2: Travel distance distributions (Real vs. Simulation).

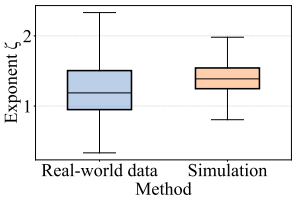

Figure 3: Visitation frequency (Real vs. Simulation).

Visitation frequency. Visitation frequency describes the rank–frequency relationship among the locations an individual visits. The visitation frequency f_k of the k-th most frequently visited location follows $f_k \propto k^{-\zeta}$, where $\zeta \approx 1.2 \pm 0.1$ in a prior study [29]. A larger value of ζ indicates that the individual concentrates visits on a small number of core locations, whereas a smaller ζ implies a greater tendency to explore new locations. For each user, we estimate

ζ and compare the resulting distributions across users between the real-world data and the simulation, as shown in Figure 3. We observe that the simulation produces insufficient heterogeneity compared with the real-world data. In the real-world dataset, some individuals tend to explore a large number of new locations, while others repeatedly return to a small set of core places. Together with the travel-distance result, this confirms that the simulation fails to reproduce population-level mobility patterns.

2.2 A preliminary study on the effectiveness of mobility scaling law guidance

In this section, we provide a preliminary validation of the effectiveness of mobility scaling-law guidance. Specifically, we focus on a single scaling-law objective, the radius of gyration. We first compute the radius-of-gyration distribution from simulated data. Next, we prompt another LLM to analyze the differences between the simulated and target radius distributions. Based on this analysis, the LLM provides coarse adjustment suggestions. We then randomly select 5% of individuals for prompt adjustment and generate new trajectories using the updated prompts. We then visualize the cumulative distribution functions (CDFs) of the radius distributions for the real-world data, the simulation, and the adjusted trajectories, as shown in Figure 4. We observe that scaling-law-guided adjustments bring the simulated trajectories closer to real-world patterns. Nevertheless, a discrepancy remains, partly because the adjustment strategy is simple and focuses on only one mobility scaling-law objective. Furthermore, scaling laws of human mobility span multiple dimensions of population-level behavior, making it unclear which specific individuals should be adjusted and whether one or multiple aspects of their behavior need to be modified.

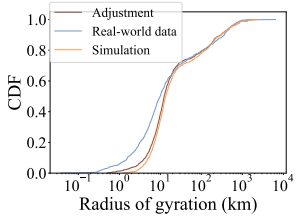

Figure 4: Comparison of radius of gyration.

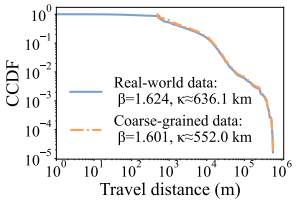

Figure 5: Travel distance (Real vs. Coarse-grained).

2.3 Preservation of human mobility scaling laws in coarse-grained data

We next ask whether coarse-grained trajectories still carry the scaling laws present in the original real-world data. Following dataset [37], we map the coordinates of the real data onto a 500 m × 500 m grid and measure all distances in this grid space; time is likewise discretized into intervals of length 48 to form coarse-grained trajectoris. On both the real and coarse-grained data, we then compare travel distance, stay duration, and visitation frequency, which together cover the spatial, temporal, and joint spatiotemporal aspects of mobility. For space, we report only the spatial result here (Figure 5) and defer the rest to Appendix A.1. In every case, the coarse-grained data preserves the same scaling laws as the full-resolution data, and this is what makes coarse and shared data usable as guidance.

3 Methodology

3.1 Definition

3.1.1 ***Trajectory.*** A mobility trajectory refers to a sequence of discrete location visits within a single day, rather than a densely sampled GPS trace. Each visit records where an individual stays at a specific time, marking a meaningful moment in their daily activities. Formally, we define the trajectory of individual i as $Y^i = \{(l_1^i, t_1^i), (l_2^i, t_2^i), \ldots, (l_{T_i}^i, t_{T_i}^i)\}$, where $l_k^i = (\text{lat}_k^i, \text{lon}_k^i) \in \mathbb{R}^2$ denotes the geographic coordinates and t_k^i denotes the continuous timestamp within the day. The trajectory length T_i is not fixed in advance and varies across individuals, reflecting heterogeneous activity intensities.

3.1.2 ***LLM-based mobility simulation model.*** We build our simulation model by adapting existing LLM-based mobility simulation approaches [1, 13]. Following these methods, we model each individual as an LLM-powered agent that generates a mobility trajectory based on a user profile, such as occupation, income, and demographic attributes. The prompt examples are provided in Appendix A.4. Formally, let $\mathcal{M}$ denote the LLM-based simulation model, and let p_i denote the prompt provided to $\mathcal{M}$ for individual i. The prompt p_i consists of three components: a user profile and a task description. The simulation process can be expressed as $Y^i = \mathcal{M}(p_i)$, where Y^i is the trajectory generated for individual i as defined in Section 3.1.1.

Extending this individual-level formulation to population-level simulation, we consider a set of n individuals, each associated with a prompt p_i. Let $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ denote the set of prompts assigned to these individuals, and let $Y = \{Y^1, Y^2, \ldots, Y^n\}$ denote the resulting trajectories generated by applying $\mathcal{M}$ to each prompt.

3.1.3 ***Problem formulation.*** Given the LLM-based simulation model $\mathcal{M}$ and a population of n individuals, whose prompts share the same task description but differ in the user profile, the goal of this work is to identify a set of personalized prompts $\mathcal{P} = \{p^1, \ldots, p^n\}$ that improve the realism of the generated population's mobility behaviors across multiple scaling laws of human mobility, by augmenting each prompt with detailed persona descriptions and behavioral constraints.

Since realism is jointly characterized by multiple scaling law metrics and each individual is assigned its own prompt, we formulate prompt adjustment for human mobility simulation as a **multi-objective, multi-prompt optimization problem**. Given the scaling law metrics of human mobility from shared data as guidance, we define a set of D target objectives $\mathcal{X}^* = \{x^{*(1)}, \ldots, x^{*(D)}\}$, where each $x^{*(d)}$ corresponds to one such scaling law metric. Our goal is to find an optimal prompt set $\mathcal{P}^*$ that jointly optimizes all objectives in $\mathcal{X}^*$.

3.2 Overview

We design COMPASS, a multi-prompt adjustment framework for LLM-based human mobility simulation, guided by mobility scaling law from shared data, as shown in Figure 6. First, we use a simulation model that takes initial prompts as input to generate a set of mobility trajectories. Second, different types of shared data provide scaling-law guidance for evaluating and improving the generated trajectories (details in Section 3.4). Third, in order to leverage the

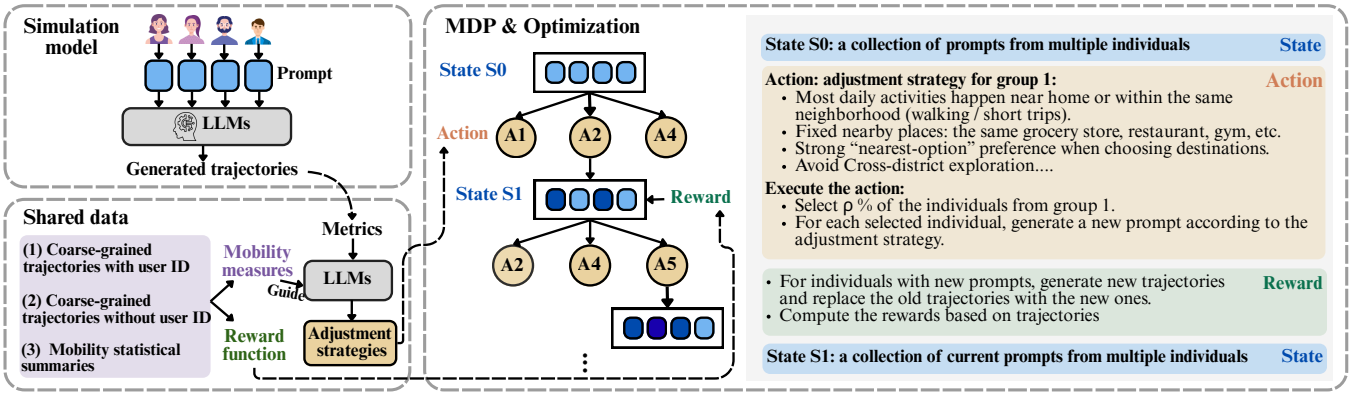

Figure 6: Framework of COMPASS.

guidance provided by the shared data to identify an optimal set of prompts for improving simulation, we formulate the prompt adjustment process as a Markov Decision Process, where the state represents the current set of prompts for all individuals, and each action corresponds to a coarse-grained prompt adjustment strategy applied to a group of individuals, generated by an LLM. We then use Monte Carlo Tree Search to find an optimal set of prompts (details in Section 3.3 and Section 3.5). Finally, this optimized set of prompts can be used for human mobility simulation and can also be extended to larger populations based on similar profiles, thus reducing the cost of prompt adjustment (details in Section 3.6).

3.3 COMPASS MDP formulation

We model the prompt adjustment process as a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r)$. $\mathcal{S}$ denotes the set of states; $\mathcal{A}$ denotes the action space; $\mathcal{T}$ denotes the state transition; r is the reward function. We introduce $\mathcal{S}, \mathcal{A}, \mathcal{T}, r$ in detail as follows.

States $\mathcal{S}$: The state $s_t \in \mathcal{S}$ represents the current set of prompts assigned to a population of n individuals:$s_t = \{p_t^1, p_t^2, \ldots, p_t^n\}$. At initialization, all individuals share the same base prompt, i.e., $p_0^i = p_{\text{init}}$, while personalization is achieved by different user profiles.

Action $\mathcal{A}$: An action $a \in \mathcal{A}$ is a group-specific prompt-adjustment strategy expressed in natural language, generated by the procedure described below:

For each shared-data as guidance channel (Sec. 3.4), we (i) compute a target mobility scaling law metric (e.g., travel distance) X^* from the shared data and the corresponding metric $\hat{X}$ from the current simulated trajectories; (ii) prompt an LLM to compare X^* with $\hat{X}$, automatically partition the population into K behaviorally coherent groups, and produce one tailored adjustment strategy $a^{(k)}$ for each group. An example prompt is provided in Appendix A.4. Each $a^{(k)}$ is treated as a single discrete action (see Figure 6 for an example). Repeating the procedure for every target scaling-law metric of human mobility (e.g., spatial and temporal) and aggregating the results yields the final action space $\mathcal{A}$. The action space is constructed once prior to MCTS and remains fixed throughout the search. Two factors decide $|\mathcal{A}|$: the number of target scaling-law metric of human mobility D (e.g., spatial, temporal) and the number of behavioral groups K the LLM identifies per measure. In our experiments this yields $|\mathcal{A}| \approx 10$–20.

Transition $\mathcal{T}$: The transition specifies how the state is updated after applying an action a_t. Each action targets a specific group of users; we sample a fraction $\rho \in (0, 1]$ of this group via a user-level bandit (details in Sec. 3.5) to form the selected user set $\mathcal{U}_t$. For each $u_i \in \mathcal{U}_t$, the new individual prompt is produced by a prompt-rewriting LLM: $p_{t+1}^i = \text{LLM}_{\text{rewrite}}(p_t^i, a_t)$. All other individuals keep their previous prompts ($p_{t+1}^i = p_t^i$). The aggregated prompts form the next state $s_{t+1} = (p_{t+1}^1, \ldots, p_{t+1}^n)$.

Reward r: The reward function r measures the quality of the simulation at each step. Let $\mathbf{x}_t = (x_t^{(1)}, \ldots, x_t^{(D)})$ denote the vector of scaling-law metrics of human mobility observed at step t, where $x_t^{(i)}$ may be either a scalar or a vector depending on the type of shared data; let $\mathbf{x}^*$ denote the corresponding target vector. For each scaling-law metric i, we define a function $g(x_t^{(i)}, x^{*(i)})$ that measures the distance to the target one, whose specific definition depends on the type of shared data (details in Section 3.4).

We normalize each distance function $g(x_t^{(i)}, x^{*(i)})$ by its initial value $g(x_0^{(i)}, x^{*(i)})$ and aggregate the D per-channel utilities via the geometric mean:

$$R(t) = \left(\prod_{i=1}^{D} \frac{g(x_0^{(i)}, x^{*(i)})}{g(x_0^{(i)}, x^{*(i)}) + g(x_t^{(i)}, x^{*(i)})} \right)^{1/D}. \tag{1}$$

The geometric mean on bounded utilities enforces balanced improvements across all dimensions [12]. The immediate reward is defined as the increase in aggregated utility between consecutive steps: $r_t = R(t+1) - R(t)$.

3.4 Scaling-law of human mobility from different types of shared data

We consider three typical types of shared data, each providing different scaling-laws of human mobility for guidance. All such scaling-laws are compatible with our framework; the key difference lies in how they influence action generation and the design of the reward function.

(1) Coarse-grained trajectories with user IDs. This type of shared data [37] is available at the user level, where each trajectory is linked to a unique but anonymous user ID. We select several scaling law metrics of human mobility as the target to guide prompt adjustment. In the following, we describe the selected target scaling law metrics and discuss the reasons for choosing them. We select the radius of gyration, stay duration, and visitation frequency as target scaling law metrics, representing the spatial, temporal, and spatial-temporal dimensions of the scaling law of human mobility,

respectively. We choose them because they are representative of the three dimensions and have been widely studied in human mobility research, where power-law scaling has often been reported [10, 29]. In addition, these metrics can be computed at the user level, enabling analysis of differences across population groups.

For action generation, we provide the LLM with the CCDF distributions of the radius of gyration and stay duration, as well as the distribution of user-specific ζ in visitation frequency, for analysis. For the reward function: we define the function $g(x_t^{(i)}, x^{*(i)})$ measuring the distance between the simulated scaling law metric $x_t^{(i)}$ and its target value $x^{*(i)}$, where $x_t^{(i)}$ is a vector. This function consists of two components: a scale-sensitive deviation from the target and a distribution discrepancy. For the first component, we compute the 1-Wasserstein distance in log space, which is defined as $\mathcal{L}_{\mathrm{W1}}(x_t^{(i)}, x^{*(i)}) = W_1\big(\log(x_t^{(i)} + \varepsilon),\ \log(x^{*(i)} + \varepsilon)\big)$, where ε is a small constant to ensure numerical stability. For the second component, we measure distance between the two CCDFs: $\mathcal{L}_{\mathrm{L1}}(x_t^{(i)}, x^{*(i)}) = \int \big|\hat{C}_t^{(i)}(z) - \hat{C}_*^{(i)}(z)\big|\,\mathrm{d}z$, where $\hat{C}_t^{(i)}$ and $\hat{C}_*^{(i)}$ denote the CCDFs of $x_t^{(i)}$ and $x^{*(i)}$. Finally, the function $g_i(x_t^{(i)}, x^{*(i)})$ is defined as $g(x_t^{(i)}, x^{*(i)}) = \mathcal{L}_{\mathrm{W1}}\Big(x_t^{(i)}, x^{*(i)}\Big) + \mu\,\mathcal{L}_{\mathrm{L1}}\Big(x_t^{(i)}, x^{*(i)}\Big)$, where μ is a weighting coefficient.

(2) Coarse-grained trajectories without user IDs. This type of shared data [42] is not available at the user level and consists only of independent trajectories. Consequently, scaling law metrics that rely on user-level information (e.g, radius of gyration) or long-term mobility history (e.g., revisit probabilities) cannot be computed. For the target scaling law metrics, we select travel distance and stay duration, which are computed at the trajectory level and represent the spatial and temporal dimensions, respectively. The travel distance is adopted as the spatial metric, as it reflects individual mobility decisions and directly influences higher-level spatial mobility metrics, such as the radius of gyration and the travel range.

For action generation, we provide the LLM with the CCDF distributions of the travel distance and stay duration, as well as the distribution of user-specific ζ in visitation frequency, for analysis. For the reward function, we use the same distance function as in the first type of shared data setting.

(3) Mobility statistical summaries This type of shared data includes only reported statistical summaries, without explicit trajectory information, e.g., parameters exponent β describing travel distance distributions from prior studies [10]. Such data provide only coarse-grained guidance for mobility simulations. For example, by comparing the exponent and cutoff parameters of travel distance distributions between real-world and simulated data, we can infer whether certain population groups are potentially overrepresented or underrepresented. Following the first type of shared data, we select the radius of gyration, stay duration, and visitation frequency as target scaling law metrics.

For action generation, the LLM is provided with statistical summaries, including the exponent and cutoff parameters of the travel distance and stay duration distributions, as well as the total ζ associated with visitation frequency. For the reward function: as the scaling law metric $x_t^{(i)}$ is a scalar, we define the function measuring the distance to the target value $x^{*(i)}$ as $g(x_t^{(i)}, x^{*(i)}) = \frac{|x_t^{(i)} - x^{*(i)}|}{x^{*(i)}}$.

3.5 Optimization

We adapt Monte Carlo Tree Search (MCTS) [34] to this MDP framework by incorporating both strategy-level and user-level selection. In our formulation, each node represents a state and each edge corresponds to a coarse-grained adjustment strategy, while an additional user-level selection step determines which individuals receive the selected strategy. A state-action value function estimates the expected cumulative reward of applying a strategy at a given state. The search tree is constructed by iteratively performing four operations: selection, expansion, simulation, and back-propagation.

Selection. At each iteration, the algorithm starts from s_0 through the search tree until reaching a leaf. Selection operates at two levels: which *action* to take and which *users* the action is applied to.

(1) Action-level: At a state s, among already-expanded actions, we select $a^* = \arg\max_a \mathrm{UCT}(s, a)$ with the score $\mathrm{UCT}(s, a) = Q(s, a) + c\sqrt{\frac{\ln(N(s)+1)}{N(s,a)+1}}$, where $Q(s, a)$ is the mean reward of (s, a), $N(s)$ and $N(s, a)$ are visit counts, and c controls exploration.

(2) User-level: Given a^*, the algorithm selects a fraction ρ of users from the group targeted by this action. Rather than sampling randomly, we maintain a per-user reward $Q_u(u_i, a, s)$ and visit count $N_u(u_i, a, s)$, and select users by

$$\mathrm{UCB}_u(u_i; a^*, s) = Q_u(u_i, a^*, s) + c_u\sqrt{\frac{\ln(N_{a^*,s} + 1)}{N_u(u_i, a^*, s) + 1}}, \qquad (2)$$

where $N_{a^*,s} = \sum_j N_u(u_j, a^*, s)$. This identifies both the most promising actions and the users most responsive to them.

Expansion. The expansion step grows the search tree by adding new child nodes under the leaf reached during selection. Evaluating all actions at every leaf is too costly, since each evaluation triggers multiple LLM calls. We therefore first filter $\mathcal{A}$ down to a small candidate set C using a *count-conditioned* global UCB, then evaluate only the candidates in C by their immediate rewards.

Since repeatedly applying the same action along a path tends to yield diminishing returns, we condition the global statistics on both the action and its in-path occurrence count. Let m_a denote the number of times action a has appeared on the path. We keep an average reward $Q_g(a, m_a)$ and visit count $N_g(a, m_a)$, and score each action by its average reward $Q_g(a, m_a)$ only when $N_g(a, m_a)$ reaches a minimum reliability threshold $n_{\min}$; otherwise, we assign it a neutral score of zero. The top-n actions under this score form the candidate set C. For each $a \in C$, the transition function yields a successor state s_a and the reward function yields an immediate reward $r(s, a)$. We select $a^* = \arg\max_{a \in C} r(s, a)$, add s_{a^*} to the tree as a new child, and forward it to the simulation stage.

Simulation. The simulation phase estimates long-term returns by performing a forward rollout starting from the newly expanded state. Given the initial state s', the algorithm iteratively applies a rollout policy, while no additional tree nodes are created during simulation. At each rollout step, we first generate a candidate action set C using the global filtering policy. For each $a \in C$, we obtain the resulting state $s_a = \mathcal{T}(s, a)$ and the immediate reward $r(s, a)$ through the transition function and the reward function. We then select the action with the highest immediate reward to advance the rollout. The rollout terminates when a predefined stopping criterion is met, such as reaching a maximum rollout depth.

Back-propagation. After a rollout completes, the observed rewards are propagated backward along the trajectory to update three sets of value estimates: the per-node $Q(s, a)$ used by UCT, the count-conditioned global $Q_g(a, m_a)$ used by candidate filtering, and the per-user $Q_u(u_i, a, s)$ used by user-level selection.

(1) Per-node update (cumulative). For each (s_t, a_t) on the trajectory, $Q(s_t, a_t)$ is the average cumulative reward of all Ω rollouts passing through it: $Q(s_t, a_t) = \frac{1}{\Omega} \sum_{j=1}^{\Omega} \sum_{(s',a') \in \text{traj}_j(s_t,a_t)} r(s', a')$, where $\text{traj}_j(s_t, a_t)$ is the sub-trajectory of rollout j starting at (s_t, a_t) until termination.

(2) Global updates. We update the global statistics with the immediate reward $r(s_t, a_t)$ instead of the cumulative reward, so that they reflect only the value contributed by (s_t, a_t) itself, without absorbing credit from subsequent actions in the trajectory.

3.6 Applying the optimized prompt set

After identifying the optimal prompts, we can directly use the updated prompt set to generate new mobility trajectories. These prompts can also be scaled to larger populations. To reduce the computational cost of searching for the optimal prompt set, we first randomly sample a subset of users and perform prompt optimization on this group. Once the optimized prompts are obtained, we scale them to the full population. Specifically, for each adjusted prompt, we assign it to the m most similar users in the full population based on profile similarity, where m is determined by the population size ratio between the subset and the full population.

4 Evaluation

4.1 Dataset description

We use two public mobility datasets. One dataset is collected in Beijing, China, covering the period from October 1, 2019 to December 31, 2019 [25]. The dataset contains mobility trajectories and user profile information (e.g., age, gender, and occupation), collected via a social networking platform. The other dataset is a global simulation dataset that has been validated under scaling laws of human mobility [42]. All trajectory points are discretized into spatial grids of 1000×1000 meters, and time is discretized into 48 time slots per day. The dataset does not contain user identifiers. We select one city (New York City) from this dataset, which includes over 100,000 trajectories. Since this dataset does not contain persistent user identifiers, we evaluate only metrics that do not require user-level long-term histories on NYC. In our simulation setting, user profile data are required as inputs, and mobility trajectories are used for validation. We select these two datasets because they satisfy both requirements and cover different countries and cities. Some other publicly available datasets are not suitable for our setting. For example, datasets [37] do not release city-level information, making it impossible to construct individual-level profiles. In addition, POI check-in datasets are less appropriate, as their sparse records do not align well with the daily mobility patterns.

For the Beijing dataset, we select 1,200 individuals with approximately one month of data as the ground truth, resulting in around 30,000 trajectories in total. All baseline methods and our approach use the released user profiles and an LLM to generate one-month synthetic human mobility trajectories. For the New York dataset, we use the entire dataset as the ground truth. We simulate user profiles based on U.S. Census demographics [5]; all baseline methods and our approach use these simulated profiles and an LLM to generate two-week synthetic human mobility trajectories.

4.2 Evaluation setup

4.2.1 ***Implementation.*** For our method COMPASS, we use the first type of shared data as scaling law guidance and implement the framework with Qwen2.5-72B as the LLM backend. All experiments are implemented in Python 3.13.5. For the transition function, we set the user sampling ratio to $\rho = 0.15$, meaning that each action is applied to 15% of users in its corresponding group. For the reward function, we set $\mu = 1$. In the MCTS search, we perform 500 simulations, and the tree depth is determined adaptively by the search process; in practice, the explored trees typically reach a depth of around 8–9. The action-level UCT exploration constant is set to $c = 1.4$. For user-level selection, we use a UCB bandit with exploration constant $c_u = 1.0$. For candidate filtering, we use count-conditioned global action values with a top-ratio of 0.6, selecting three candidate actions at each non-root expansion. Following Section 3.6, for the Beijing dataset, we perform prompt optimization on 30% of users to reduce search cost, and then extend the optimized prompts to the full population. Specifically, for each optimized prompt, we assign it to the most similar users in the full population based on profile similarity, which is computed using the pre-trained MPNet model [30].

4.2.2 ***Baselines.*** We compare our method with LLM-based human mobility simulation baselines: **CoPB [25]** is a mobility simulation framework that guides LLMs through structured reasoning stages to generate realistic mobility intentions. **Urban-Mobility-LLM (UML) [1]** is a method that synthesizes travel survey data by prompting LLMs to generate individual mobility patterns. **LLMob [13]** is an LLM-based agent framework for personal mobility generation, combining self-consistency and retrieval strategies. **CitySim [2]** leverages LLM-powered agents with personas, memory, and long-term goals to simulate realistic human behavior.

We also conduct comparisons with several variants of the model: (1) We consider two additional types of shared data that provide scaling law guidance, namely **COMPASS with the second Shared Data (w SD2)** and **COMPASS with the third Shared Data (w SD3)**. (2) **COMPASS w/o PA**. In this setting, we remove the prompt adjustment part. (3) **COMPASS w E**. In this setting, we perform prompt search on a sampled subset and extend the resulting adjustments to the full population based on profile similarity. (4) **COMPASS w/o UB**. This variant replaces the per-user UCB bandit with random user sampling when selecting individuals for a chosen adjustment strategy. (5) **COMPASS w/o UB & CUCB**. This variant further replaces the count-conditioned global UCB for candidate filtering with a standard action-level UCB that does not distinguish between different application counts of the same adjustment.

4.2.3 ***Metric.*** Following existing simulation work [13, 25], we evaluate our model using three aspects metrics. To assess the similarity between the simulation and real-world data, we compute the Jensen–Shannon divergence (JSD) between their distributions for each metric. It is worth noting that the evaluation metrics include the scaling law metrics used as guidance. While we report results

Table 1: Overall Performance on Beijing dataset. Bold scores are for the best values. The gray-shaded parts indicate the metrics used for guidance under the specific shared data setting, while the remaining metrics are used for evaluation.

Method	Spatial			Temporal		Spatial-temporal		
	Radius	Distance	Locfreq	Duration	Circadian	Visit-Freq	Exploration	Return
CoPB	0.0663±0.0093	0.0283±0.0014	0.2548±0.0630	0.0309±0.0008	0.1229±0.0336	0.1471±0.0039	0.1445±0.0275	0.2436±0.0146
UML	0.0565±0.0037	0.0291±0.0078	0.2711±0.0173	0.0234±0.0026	0.0913±0.0020	0.1752±0.0406	0.1829±0.0399	0.2669±0.0735
LLMob	0.0963±0.0048	0.0096±0.0016	0.3111±0.0083	0.0305±0.0026	0.0906±0.0005	0.1607±0.0337	0.3895±0.0290	0.2531±0.0383
CitySim	0.0584±0.0042	0.0531±0.0044	0.2616±0.0005	0.0325±0.0018	0.1129±0.0081	0.1507±0.0181	0.1285±0.0672	0.3452±0.0115
COMPASS w SD2	0.0376±0.0014	**0.0065±0.0009**	0.1918±0.0019	0.0178±0.0003	0.0735±0.0002	0.1315±0.0020	0.1001±0.0069	0.0803±0.0009
COMPASS w SD3	0.0403±0.0024	0.0069±0.0004	0.1957±0.0058	**0.0174±0.0005**	0.0732±0.0003	0.1391±0.0043	0.0854±0.0018	0.0768±0.0023
COMPASS	**0.0324±0.0021**	0.0070±0.0004	**0.1914±0.0016**	0.0180±0.0004	**0.0718±0.0007**	**0.1247±0.0032**	**0.0756±0.0134**	**0.0614±0.0034**

Table 2: Overall Performance on NYC dataset. Bold scores are for the best values. The gray-shaded parts indicate the metrics used for guidance under the specific shared data setting, while the remaining metrics are used for evaluation.

Method	Spatial		Temporal	
	Distance	Locfreq	Duration	Circadian
CoPB	0.1370±0.0126	0.2893±0.0155	0.1326±0.0301	0.1046±0.0436
UML	0.1097±0.0026	0.2812±0.0339	0.1444±0.0016	0.0666±0.0069
LLMob	0.1395±0.0048	0.3451±0.0048	0.1400±0.0005	0.0745±0.0059
CitySim	0.1359±0.0062	0.2912±0.0107	0.0869±0.0037	0.1620±0.0237
COMPASS w SD2	0.0397±0.0005	**0.2511±0.0503**	**0.0721±0.0002**	**0.0547±0.0143**
COMPASS w SD3	0.0411±0.0009	0.2649 ±0.0283	0.0854±0.0005	0.0578 ± 0.0156

for all metrics, performance improvements are evaluated only on metrics excluding the guidance metrics.

Spatial aspect: (1) Radius of gyration: quantifies the spatial extent of an individual's mobility. We compute the radius of gyration as the root mean square distance of visited locations from the trajectory centroid. (2) Travel distance: defined as the geographical distance between consecutive locations in a trajectory. (3) Origin–destination similarity (OdSim): quantifies the similarity between generated and real trajectories in terms of OD travel patterns. It is computed by comparing the normalized frequency distributions of OD pairs using JSD.

Temporal aspect: (1) Stay duration: measures how long individuals remain at each visited location. We compute the duration distribution from the time intervals between consecutive movements. (2) Circadian rhythm: This metric [24] captures the circadian rhythm of human mobility intensity over a 24-hour period. We quantify it by aggregating trips into hourly bins.

Spatial-temporal aspect: (1) Visitation frequency. This metric [29] describes the rank-frequency relationship of an individual's visited locations. The detailed definition is in Section 2. (2) Exploration: This metric [29] characterizes how the number of distinct locations grows with the total number of visits. Let H denote the number of stays and $N_{\text{new}}(H)$ the cumulative number of unique locations. Exploration follows a sublinear power-law, $N_{\text{new}}(H) \sim H^{\alpha}$, where α is the exploration exponent. We estimate α for each user and compare its distribution between real and simulated data. (3) Preferential return: Following [29], we characterize preferential return as the tendency to revisit previously visited locations with probability proportional to their past visitation frequency. Let $n_i(t)$ denote the number of visits to location i prior to time t. Following the original linear assumption ($P \propto n_i$), where P denotes the probability of returning to location i, we generalize the model by fitting $P \propto n_i^{\gamma}$. We estimate γ for each user and compare the simulated and real user-level γ distributions using JSD.

4.3 Overall performance

4.3.1 ***Overall performance on Beijing dataset.*** We compare our method COMPASS with other baseline methods, as shown in the Table 1. For our method COMPASS, we use the first type of shared data to provide guidance. As guidance, we use three scaling law metrics of human mobility: radius of gyration, stay duration, and visitation frequency, corresponding to the spatial, temporal, and spatio-temporal perspectives, respectively. The guidance metrics are shown in gray, while the remaining metrics are used only for evaluation. From the Table, we can see that our method outperforms all other baselines, achieving at least an improvement of 27.08% in travel distance, 24.88% in odSim, 20.75% in circadian, 41.17% in exploration, and 74.79% in return. This demonstrates that scaling laws of human mobility can provide effective guidance for individual-level generation, thereby improving collective realism.

4.3.2 ***Overall performance on NYC dataset.*** For the NYC dataset, we only consider the latter two types of shared data and evaluate spatial and temporal metrics, since the dataset does not provide user-level trajectories and thus does not support spatial-temporal metric computation. As guidance, we use two scaling law metrics of human mobility: radius of gyration and stay duration, corresponding to the spatial and temporal perspectives, respectively. The guidance metrics are shown in gray, while the remaining metrics are used only for evaluation. We compare our method COMPASS with other baseline methods, as shown in Table 2. From the Table, we can see that our method outperforms all other baselines, achieving at least an improvement of 10.70% in OdSim, 17.87% in Circadian.

4.4 Performance across shared data types

In this section, we study how different types of shared data affect prompt adjustment for simulation, and whether good performance can still be achieved using only statistical summaries.

4.4.1 ***Performance across shared data types on Beijing dataset.*** For the Beijing dataset, we consider three types of shared data and select metrics from three perspectives: spatial, temporal, and spatio-temporal. We compare our method COMPASS with variants, as shown in the Table 1. Due to data type limitations, different shared data settings use different scaling laws of human mobility as guidance. Therefore, we use the metrics not included as guidance for evaluation, as shown in gray. We can observe that using the

first type of shared data (i.e., COMPASS) achieves the best performance, because each scaling law metric is computed at the user level, allowing more precise identification of issues across different individuals. The performance of using the third type of shared data (i.e., COMPASS w SD3) slightly decreases, since having only mobility statistical summaries can provide only a coarse direction for improvement. Nevertheless, the overall performance of COMPASS w SD3 remains good and outperforms all baselines.

4.4.2 ***Performance across shared data types on NYC dataset.*** For the New York City dataset, we consider two types of shared data and select metrics from spatial and temporal aspects. We compare two variants of our method COMPASS, as shown in Table 2. Similarly to the Beijing dataset, we use the metrics not included as the guidance for evaluation, as shown in gray. Overall, using the second shared data (i.e., COMPASS w SD2) performs slightly better than using the third shared data (i.e., COMPASS w SD3), because COMPASS w SD2 provides more fine-grained rewards. Similar to the findings in the Beijing dataset, even when COMPASS w SD3 uses mobility statistical summaries as the guidance, performance remains good and outperforms all baselines.

4.5 Effect of prompt adjustment on mobility scaling laws

To assess whether scaling-law-guided prompt adjustment improves behavioral realism, we compare simulations before (COMPASS w/o PA) and after adjustment (COMPASS) on the Beijing dataset under the first shared-data setting, using well-established scaling laws of human mobility: travel distance, preferential return, and exploration scaling. These analyses complement the metric-based evaluation by examining whether the generated trajectories reproduce key behavioral patterns observed in real-world data.

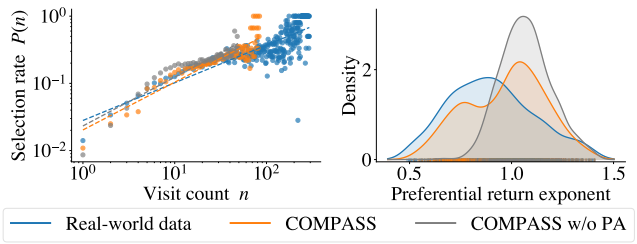

Figure 7: Preferential return behavior.

Preferential return: Figure 7 examines preferential return behavior from two perspectives. The left panel shows the relationship between a location's visit count n and its selection rate $P(n)$. Compared with w/o PA, COMPASS more closely matches the real-data trend, particularly for highly visited locations. The right panel compares the distribution of the preferential-return exponent γ estimated separately for each user. COMPASS produces a distribution that is substantially closer to the real data, while w/o PA shifts toward larger γ values. This suggests that our framework not only improves the aggregate preferential-return pattern but also better captures user-level heterogeneity in revisit behavior.

Exploration: Figure 8 examines exploration behavior. The left panel shows the growth of explored locations over time. Although COMPASS slightly overestimates exploration compared with the real data, it substantially reduces the gap observed in w/o PA. The right panel compares the distribution of the user-level exploration exponent α, where COMPASS more closely matches the real distribution. These results suggest that our framework improves the modeling of exploration dynamics at both the population and individual levels.

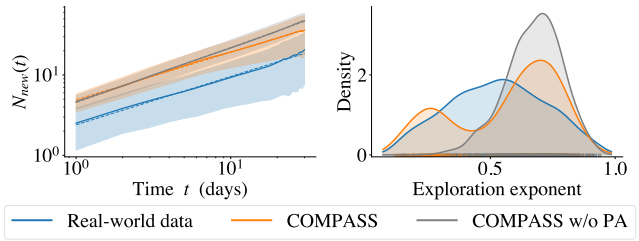

Figure 8: Exploration scaling behavior.

Travel distance: Figure 9 shows the results for travel distance. For visualization clarity, travel distances are clipped to a predefined maximum threshold (i.e., 600km) during plotting; all reported metrics are computed using the full dataset. We observe that, after adjustment, the travel distance distribution becomes closer to the real-world data. In addition, the cutoff values obtained from fitting the truncated power-law distribution are more consistent with those of the real-world data, better reflecting realistic human activity ranges.

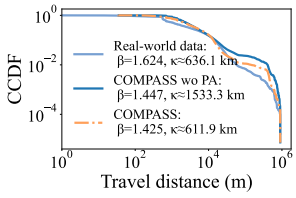

Figure 9: Effect of prompt adjustment on travel distance.

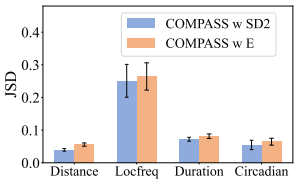

Figure 10: Performance of optimal prompt scaling from a subset to the full population.

4.6 Ablation study

4.6.1 Subset-search generalization. To validate whether the prompt adjustments identified from a small subset can generalize to a larger population, we conduct an experiment using the NYC dataset. We choose this dataset because it contains fewer users and trajectories, making full-dataset search computationally feasible. Specifically, we compare COMPASS w E, which performs prompt search on a subset and extends the resulting adjustments to the full population, with COMPASS w SD2, which performs prompt search directly on the full dataset. Figure 10 shows the two variants achieve comparable performance, indicating that prompt adjustments learned from a subset can effectively generalize to larger groups of individuals.

4.6.2 ***Effect of the user-bandit and count-conditioned UCB.*** To assess the contribution of the two key components in COMPASS, we compare the full model with two variants on the Beijing dataset under the first shared-data setting. We report performance on the evaluation metrics that are not directly used as scaling law guidance. COMPASS w/o UB replaces the per-user UCB bandit with random user sampling when selecting users for an action. COMPASS w/o UB & CUCB further replaces the count-conditioned global UCB

with a standard action-level UCB that does not distinguish between different application counts of the same adjustment.

Figure 11 reports the performance of the full model and its variants on the five evaluation metrics. Removing the user-bandit consistently worsens performance across all metrics, while removing both the user-bandit and count-conditioned UCB leads to a further decline. These results confirm that both components contribute to the effectiveness of COMPASS.

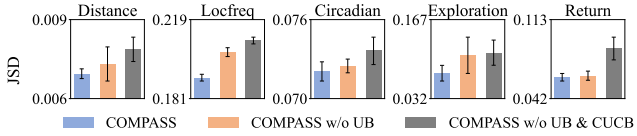

Figure 11: Effect of the user-bandit and count-conditioned UCB.

4.6.3 ***Effect of MCTS-based optimization.*** To evaluate the contribution of the proposed MCTS-based search, we compare COMPASS with a random-search variant (COMPASS w/ RS) on the Beijing dataset under the first shared-data setting. This variant uses the same action space as COMPASS but selects actions uniformly at random from the candidate set, without considering future rewards.

Figure 12 shows that replacing MCTS with random search substantially degrades performance across all metrics. The largest performance drops are observed on Exploration and Return, suggesting that random action selection struggles to capture realistic mobility dynamics. These results highlight the importance of MCTS-based search. At each optimization step, the framework must determine both what adjustment to apply and which users should receive it. Random search makes these decisions without considering their estimated utility, whereas MCTS explicitly evaluates candidate actions before selecting them. As a result, MCTS is more likely to identify adjustments that improve population-level mobility statistics, leading to consistently lower discrepancies across all metrics.

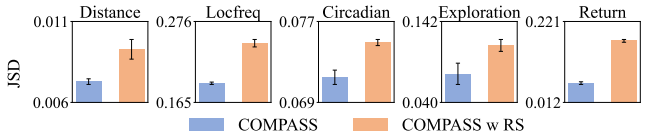

Figure 12: Comparison between COMPASS and random search.

4.7 Performance across different LLMs

To examine whether COMPASS depends on a particular LLM backbone, we evaluate it on the Beijing dataset under the first shared-data setting using four representative models: GPT-5.2, GPT-4o-mini, LLaMA 3.3 70B, and Qwen 2.5 72B.

Table 3 reports the performance of COMPASS under four different LLM backbones across mobility metrics. We find that no single model consistently performs best across all metrics. For example, GPT-5.2 achieves the lowest discrepancy on some metrics, while Qwen 72B performs best on others. More importantly, the differences among the stronger models are relatively small. In particular, Qwen 72B matches GPT-5.2 by achieving the best result on four of the eight metrics. This suggests that the effectiveness of COMPASS does not rely on a specific LLM backbone. Open-source models can serve as a practical alternative to proprietary ones without substantially sacrificing simulation quality.

4.8 Parameter sensitivity analysis

The sampling rate k controls the fraction of users selected for each action. A smaller k reduces the number of LLM calls but provides less feedback for evaluating candidate adjustments, while a larger k increases computational cost by applying adjustments to more users. We vary $k \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ on the Beijing dataset under the first shared-data setting and report the performance on its five corresponding evaluation metrics in Figure 13.

Figure 13 shows that $k = 0.15$ achieves the best overall performance, yielding the best results on four of the five evaluation metrics. Both smaller and larger values of k lead to worse results, suggesting that an intermediate sampling rate provides the best balance between evaluation quality and adjustment granularity.

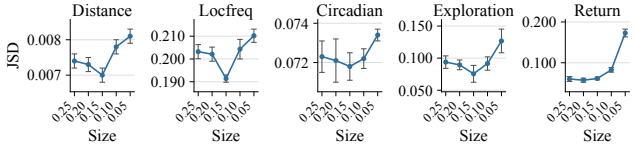

Figure 13: Sensitivity of COMPASS to the sampling rate.

4.9 Convergence analysis

To examine whether the MCTS optimization converges within a fixed number of iterations, we track the best Q value during the search on the Beijing dataset under the first shared-data setting. Figure 14 shows the rolling mean (± std), computed using a 10-iteration window. The COMPASS curves begin to plateau after roughly 500 iterations, with only marginal improvements thereafter. Therefore, we set 500 iterations as the default optimization budget for all experiments.

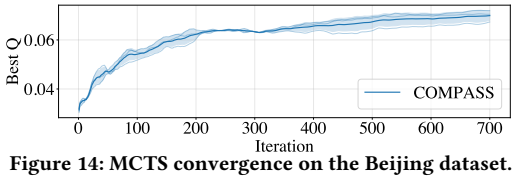

Figure 14: MCTS convergence on the Beijing dataset.

4.10 Overhead analysis

We quantify the overhead using two metrics: the number of LLM calls and the monetary cost. In our implementation, one LLM call refers to either one prompt-adjustment call that updates a user's behavioral traits, or one trajectory-generation call that generates a one-day trajectory for a user. Thus, the number of LLM calls measures the amount of fresh generation required during search, while the cost is estimated based on the provider's token pricing.

To reduce repeated LLM calls, we use a user-level cache. Each time an action is executed, we cache the before-and-after adjustment state of the selected users together with their regenerated trajectories. If the same action is later applied under the same state and selects a user already stored in the cache, we directly reuse the cached trajectory instead of invoking the LLM. In 500 MCTS iterations, the cache achieved a 95.9% hit rate, reducing the search overhead to an average of 31.5 LLM calls per iteration. Under Qwen2.5 pricing, this corresponds to a monetary cost of $5.97.

Table 3: Performance across different LLMs. Bold scores are for the best values. The gray-shaded parts indicate the metrics used for guidance under the specific shared data setting, while the remaining metrics are used for evaluation.

Method	Spatial			Temporal		Spatial-temporal		
	Radius	Distance	Locfreq	Duration	Circadian	Visit-Freq	Exploration	Return
COMPASS w GPT5.2	0.0385±0.0026	**0.0064±0.0005**	0.2017±0.0076	0.0221±0.0002	0.0738±0.0003	**0.1095±0.0042**	**0.0645±0.0014**	**0.0554±0.0101**
COMPASS w LLama3.3	0.0403±0.0008	0.0068±0.0001	0.2081±0.0018	0.0223±0.0002	0.0731±0.0002	0.1746±0.0093	0.0618±0.0030	0.0745±0.0018
COMPASS w GPT4o-mini	0.0410±0.0435	0.0077±0.0010	0.2274±0.0106	0.0229±0.0011	0.0721±0.0007	0.1612±0.0072	0.0751±0.0239	0.1129±0.0606
COMPASS w Qwen	**0.0324±0.0011**	0.0070±0.0002	**0.1914±0.0016**	**0.0180±0.0004**	**0.0718±0.0007**	0.1247±0.0032	0.0756±0.0134	0.0614±0.0034

5 Discussion

Lessons learned: Based on the results of our paper, we summarize the following lessons learned: (1) Scaling laws of human mobility obtained from different types of shared data can effectively guide LLM-based mobility simulations, as shown in Table 1 and Table 2. Even statistical information without any trajectories leads to noticeable performance improvements. (2) Our framework incorporates multi-dimensional scaling law metrics as guidance, enabling improvements that extend beyond targeted aspects and enhance the overall mobility simulation, as shown in Table 1. (3) Our framework enables identifying optimal prompt combinations from a small subset of individuals and generalizing them to a larger population, reducing search overhead, as shown in Figure 10.

Limitation: Despite its effectiveness, our framework has several limitations. Our framework introduces additional computational overhead in mobility simulation, particularly at large scales, although this cost is partially mitigated by our method. Our evaluation is currently limited to two cities and extending validation to more diverse urban environments remains future work.

Ethics and privacy: This work focuses on population-level mobility simulation rather than individual tracking or prediction. Our framework uses user profiles as inputs, which are derived from census statistics or publicly available data. All information is highly anonymized and represented at a coarse-grained level, preventing identification of individuals. Scaling laws of human mobility are obtained from publicly shared data and do not involve privacy-sensitive information. While LLM-based simulations may introduce biases or unrealistic behaviors, our framework uses scaling laws of mobility to guide individual generation for improving realism.

6 Related work

LLM-based human mobility simulation. Existing works explored using LLMs for human mobility simulation by leveraging their knowledge and human-like reasoning capabilities. The advantage of these approaches is that they do not require large-scale real-world mobility data for training. These works [1, 8, 13, 14, 16, 19, 20, 22, 25] guide LLMs to simulate human-like mobility intention reasoning step by step and then produce realistic mobility activity sequences. For example, CoPB [25] is an intention- and planning-based framework that enables LLMs to generate human mobility trajectories through step-by-step reasoning. Although these works produce realistic outputs, the reliance on LLMs makes the simulation expensive. This work [13] designs a LLM-based agent framework for personal mobility generation, combining self-consistency and retrieval strategies to align language models with real-world human activity. However, they generate each individual's trajectory independently, without any population-level coordination mechanism, thereby failing to capture the emergence of collective behaviors.

Prompt optimization Previous research on prompt optimization focuses on identifying an optimal prompt that enables an LLM to produce the best performance for a given class of tasks [7, 32, 35]. Existing methods can be categorized into three types. The first category focuses on how to select the optimal prompt given available candidates [11, 26]. For example, given a set of human-readable prompt candidates with unknown performance, ZOPO [11] embeds prompts into a continuous space and estimates update directions based on past performance, enabling efficient local exploration toward high-performing prompts. The second category considers both prompt refinement and optimal prompt selection, and addresses single-objective optimization [34, 40]. For example, Promptagent [34] enables an LLM to iteratively refine prompts based on observed errors, using MCTS to explore the space of prompt modifications. The third category considers both prompt refinement and prompt selection, with a focus on multi-objective optimization [12, 27, 28, 41, 43]. These works focus on balancing multiple objectives. For example, ParetoPrompt [43] uses a pre-trained language model as a policy, which is further optimized via reinforcement learning to achieve multi-objective prompt optimization. Unlike prior work that focuses on finding a single prompt to achieve good performance on a specific task (single or multiple objectives), we aim to address a multi-prompt, multi-objective optimization problem. Specifically, our goal is to find an optimal set of prompts that jointly optimizes multiple objectives.

7 Conclusion

In this work, we design COMPASS for LLM-based mobility simulation. It leverages scaling laws of human mobility from shared data as guidance to adjust individual-level prompts in order to reproduce realistic population-level mobility behavior. Our framework applies coarse-grained adjustment strategies guided by scaling laws of human mobility, progressively enabling fine-grained individual-level adaptation while satisfying multiple population-level mobility objectives under a limited budget. Experiments show that COMPASS significantly outperforms other LLM-based simulations.

References

[1] Prabin Bhandari, Antonios Anastasopoulos, and Dieter Pfoser. 2024. Urban mobility assessment using llms. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*. 67–79.

[2] Nicolas Bougie and Narimawa Watanabe. 2025. 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*. 215–229.

[3] Stacey Bricka, Timothy Reuscher, Paul Schroeder, Mitchell Fisher, Justina Beard, and Xiaoyuan Layla Sun. 2024. Summary of travel trends: 2022 national household travel survey. (2024).

[4] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. 2006. The scaling laws of human travel. *Nature* 439, 7075 (2006), 462–465.

[5] U.S. Census Bureau. 2020. American Community Survey (ACS). https://www.census.gov/programs-surveys/acs.html.

[6] Serina Chang, Mandy L Wilson, Bryan Lewis, Zakaria Mehrab, Komal K Dudakiya, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, et al. 2021. Supporting covid-19 policy response with large-scale mobility-based modeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2632–2642.

[7] Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024. Black-box prompt optimization: Aligning large language models without model training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3201–3219.

[8] Yuwei Du, Jie Feng, Jian Yuan, and Yong Li. 2025. CAMS: A CityGPT-Powered Agentic Framework for Urban Human Mobility Simulation. *arXiv preprint arXiv:2506.13599* (2025).

[9] Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. 2020. Learning to simulate human mobility. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 3426–3433.

[10] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779–782.

[11] Wenyang Hu, Yao Shu, Zongmin Yu, Zhaoxuan Wu, Xiaoqiang Lin, Zhongxiang Dai, See-Kiong Ng, and Bryan Kian Hsiang Low. 2024. Localized zeroth-order prompt optimization. *Advances in Neural Information Processing Systems* 37 (2024), 86309–86345.

[12] Yasaman Jafari, Dheeraj Mekala, Rose Yu, and Taylor Berg-Kirkpatrick. 2024. Morl-prompt: An empirical analysis of multi-objective reinforcement learning for discrete prompt optimization. *arXiv preprint arXiv:2402.11711* (2024).

[13] WANG JIAWEI, Renhe Jiang, Chuang Yang, Zengqing Wu, Ryosuke Shibasaki, Noboru Koshizuka, Chuan Xiao, et al. 2024. Large language models as urban residents: An llm agent framework for personal mobility generation. *Advances in Neural Information Processing Systems* 37 (2024), 124547–124574.

[14] Chenlu Ju, Jiaxin Liu, Shobhit Sinha, Hao Xue, and Flora Salim. 2025. Trajllm: A modular llm-enhanced agent-based framework for realistic human trajectory simulation. In *Companion Proceedings of the ACM on Web Conference 2025*. 2847–2850.

[15] Pierre-Yves Lajoie, Bobak Hamed Baghi, Sachini Herath, Francois Hogan, Xue Liu, and Gregory Dudek. 2024. PEOPLEx: Pedestrian opportunistic positioning leveraging IMU, UWB, BLE and WiFi. In *ICC 2024-IEEE International Conference on Communications*. IEEE, 3518–3523.

[16] Siyu Li, Toan Tran, Haowen Lin, John Krumm, Cyrus Shahabi, Lingyi Zhao, Khurram Shafique, and Li Xiong. 2024. Geo-llama: Leveraging llms for human mobility trajectory generation with spatiotemporal constraints. *arXiv preprint arXiv:2408.13918* (2024).

[17] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926* (2017).

[18] Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. 2023. Largest: A benchmark dataset for large-scale traffic forecasting. *Advances in Neural Information Processing Systems* 36 (2023), 75354–75371.

[19] Yifan Liu, Xishun Liao, Haoxuan Ma, Brian Yueshuai He, Chris Stanford, and Jiaqi Ma. 2024. Human Mobility Modeling with Household Coordination Activities under Limited Information via Retrieval-Augmented LLMs. *arXiv preprint arXiv:2409.17495* (2024).

[20] Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. 2024. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563* (2024).

[21] Kun Ouyang, Reza Shokri, David S Rosenblum, and Wenzhuo Yang. 2018. A non-parametric generative model for human trajectories.. In *IJCAI*, Vol. 18. 3812–3817.

[22] Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691* (2025).

[23] Markus Schläpfer, Lei Dong, Kevin O'Keeffe, Paolo Santi, Michael Szell, Hadrien Salat, Samuel Anklesaria, Mohammad Vazifeh, Carlo Ratti, and Geoffrey B West. 2021. The universal visitation law of human mobility. *Nature* 593, 7860 (2021), 522–527.

[24] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. 2013. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface* 10, 84 (2013), 20130246.

[25] Chenyang Shao, Fengli Xu, Bingbing Fan, Jingtao Ding, Yuan Yuan, Meng Wang, and Yong Li. 2024. Chain-of-planned-behaviour workflow elicits few-shot mobility generation in llms. *arXiv preprint arXiv:2402.09836* (2024).

[26] Chengshuai Shi, Kun Yang, Zihan Chen, Jundong Li, Jing Yang, and Cong Shen. 2024. Efficient prompt optimization through the lens of best arm identification. *Advances in Neural Information Processing Systems* 37 (2024), 99646–99685.

[27] Somanshu Singla, Zhen Wang, Tianyang Liu, Abdullah Ashfaq, Zhiting Hu, and Eric Xing. 2024. Dynamic rewarding with prompt optimization enables tuning-free self-alignment of language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 21889–21909.

[28] Ankita Sinha, Wendi Cui, Kamalika Das, and Jiaxin Zhang. 2024. Survival of the Safest: Towards Secure Prompt Optimization through Interleaved Multi-Objective Evolution. *arXiv preprint arXiv:2410.09652* (2024).

[29] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. 2010. Modelling the scaling properties of human mobility. *Nature physics* 6, 10 (2010), 818–823.

[30] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems* 33 (2020), 16857–16867.

[31] Eran Toch, Boaz Lerner, Eyal Ben-Zion, and Irad Ben-Gal. 2019. Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems* 58, 3 (2019), 501–523.

[32] Prashant Trivedi, Souradip Chakraborty, Avinash Reddy, Vaneet Aggarwal, Amrit Singh Bedi, and George K Atia. 2025. Align-pro: A principled approach to prompt optimization for llm alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 27653–27661.

[33] Jingwei Wang, Qianyue Hao, Wenzhen Huang, Xiaochen Fan, Qin Zhang, Zhentao Tang, Bin Wang, Jianye Hao, and Yong Li. 2025. Coopride: Cooperate all grids in city-scale ride-hailing dispatching with multi-agent reinforcement learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 1457–1468.

[34] Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. 2023. Promptagent: Strategic planning with language models enables expert-level prompt optimization. *arXiv preprint arXiv:2310.16427* (2023).

[35] Yurong Wu, Yan Gao, Bin Benjamin Zhu, Zineng Zhou, Xiaodi Sun, Sheng Yang, Jian-Guang Lou, Zhiming Ding, and Linjun Yang. 2024. Strago: Harnessing strategic guidance for prompt optimization. *arXiv preprint arXiv:2410.08601* (2024).

[36] Feng Xie, Zhong Zhang, Liang Li, Bin Zhou, and Yusong Tan. 2022. EpiGNN: Exploring spatial transmission with graph neural network for regional epidemic forecasting. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 469–485.

[37] Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, Yoshihide Sekimoto, Kaoru Sezaki, Esteban Moro, and Alex Pentland. 2024. YJMob100K: City-scale and longitudinal dataset of anonymized human mobility trajectories. *Scientific Data* 11, 1 (2024), 397.

[38] Xiaodong Yan, Tengwei Song, Yifeng Jiao, Jianshan He, Jiaotuan Wang, Ruopeng Li, and Wei Chu. 2023. Spatio-temporal hypergraph learning for next POI recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*. 403–412.

[39] Song Yang, Jiamou Liu, and Kaiqi Zhao. 2022. Getnext: trajectory flow map enhanced transformer for next poi recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on research and development in information retrieval*. 1144–1153.

[40] Sheng Yang, Yurong Wu, Yan Gao, Zineng Zhou, Bin Benjamin Zhu, Xiaodi Sun, Jian-Guang Lou, Zhiming Ding, Anbang Hu, Yuan Fang, et al. 2024. Ampo: Automatic multi-branched prompt optimization. *arXiv preprint arXiv:2410.08696* (2024).

[41] Haoqi Yuan, Yuhui Fu, Feiyang Xie, and Zongqing Lu. 2024. Pre-trained multi-goal transformers with prompt optimization for efficient online adaptation. *Advances in Neural Information Processing Systems* 37 (2024), 55086–55114.

[42] Yuan Yuan, Yuheng Zhang, Jingtao Ding, and Yong Li. 2025. WorldMove, a global open data for human mobility. *arXiv preprint arXiv:2504.10506* (2025).

[43] Guang Zhao, Byung-Jun Yoon, Gilchan Park, Shantenu Jha, Shinjae Yoo, and Xiaoning Qian. 2025. Pareto prompt optimization. In *The Thirteenth International Conference on Learning Representations*.

[44] Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James Yu. 2023. Difftraj: Generating gps trajectory with diffusion probabilistic model. *Advances in Neural Information Processing Systems* 36 (2023), 65168–65188.

A Appendix

A.1 Supplementary analysis: preservation of human mobility scaling laws in coarse-grained data

In this section, we provide additional results that complement Section 2.3. Based on both real-world and coarse-grained data, we compare stay duration distributions and visitation frequency from temporal and spatial-temporal perspectives (Figures 15 and 16). We find that the distributions obtained from coarse-grained data closely match those obtained from the original real-world trajectories, indicating that coarse-graining preserves key scaling laws of mobility.

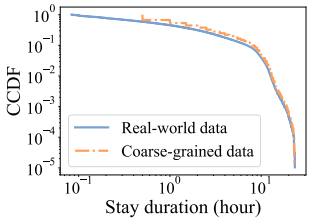

Figure 15: Stay duration distributions (Real vs. Coarse-grained).

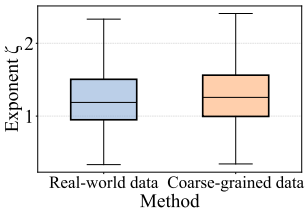

Figure 16: Visitation frequency (Real vs. Coarse-grained).

A.2 COMPASS optimization algorithm

To make the optimization process easier to follow, we summarize the main steps of COMPASS in Algorithm 1.

A.3 Example of prompt adjustment

To illustrate how behavioral constraints are injected into user profiles, we present two examples in which an initial profile is augmented with an additional behavioral description. The original profile contains only basic attributes, such as age, occupation, education level, consumption level, and coarse home/work locations. The adjusted profile further specifies mobility preferences that guide the generation model toward a particular behavioral pattern. For conciseness, the examples below report only the profile field. All other parts of the prompt, including the task description, formatting instruction remain fixed.

Example 1: prompt adjustment

Initial profile: A 35-year-old male logistics worker in Beijing with a high-school education level and medium consumption level. His home is near [116.5, 39.8], and his workplace is near [116.7, 38.0].

Profile after adjustment: [Initial profile]. In addition, the user has very weak location preference and high variability in visited places. His movements are largely driven by tasks, assignments, or service-related needs rather than routine. There is no single dominant activity center, and his mobility naturally spans multiple areas.

Algorithm 1 COMPASS optimization algorithm

Input: initial prompt state s_0, action space $\mathcal{A}$, target objectives $\mathbf{x}^*$, user sampling ratio ρ, number of iterations N_{iter}
Initialize root node v_0 with state s_0
for $i = 1$ to N_{iter} **do**
  Initialize empty trajectory τ
  ▷ Selection
  while v has expanded children and v is not terminal **do**
    $a \leftarrow \textsc{SelectActionByUCT}(v)$
    $\mathcal{U} \leftarrow \textsc{SelectUsersByUCB}(v.state, a, \rho)$
    Append $(v.state, a, \mathcal{U})$ to τ
    $v \leftarrow \text{child}(v, a)$
  end while
  ▷ Expansion
  if v is not terminal **then**
    $C \leftarrow \textsc{GlobalFilter}(v.state, \mathcal{A}, B)$
    for all $a \in C$ **do**
      $\mathcal{U}_a \leftarrow \textsc{SelectUsersByUCB}(v.state, a, \rho)$
      $(r_a, s_a) \leftarrow \textsc{ApplyAndEvaluate}(v.state, a, \mathcal{U}_a, \mathbf{x}^*)$
    end for
    $a^* \leftarrow \arg\max_{a \in C} r_a$
    Create child node v' with state s_{a^*}
    Add edge (v, a^*) to the search tree
    Append $(v.state, a^*, \mathcal{U}_{a^*}, r_{a^*})$ to τ
    $v \leftarrow v'$
  end if
  ▷ Simulation
  $s_{\text{roll}} \leftarrow v.state$
  for $t = 1$ to L **do**
    if s_{roll} is terminal **then**
      break
    end if
    $C \leftarrow \textsc{GlobalFilter}(s_{\text{roll}}, \mathcal{A}, B)$
    for all $a \in C$ **do**
      $\mathcal{U}_a \leftarrow \textsc{SelectUsersByUCB}(s_{\text{roll}}, a, \rho)$
      $(r_a, s_a) \leftarrow \textsc{ApplyAndEvaluate}(s_{\text{roll}}, a, \mathcal{U}_a, \mathbf{x}^*)$
    end for
    $a^* \leftarrow \arg\max_{a \in C} r_a$
    Append $(s_{\text{roll}}, a^*, \mathcal{U}_{a^*}, r_{a^*})$ to τ
    $s_{\text{roll}} \leftarrow s_{a^*}$
  end for
  ▷ Back-propagation
  Update node-level statistics with cumulative returns
  Update global action statistics with immediate rewards
  Update user-level statistics for selected users
end for
return the optimized prompt state from the search tree

Example 2: prompt adjustment

Initial profile: A 23-year-old female IT engineer in Beijing with a bachelor's degree and medium consumption level. Her home is near [116.3, 40.0], and her workplace is near [116.3, 40.0].

Profile after adjustment: [Initial profile]. In addition, most of her activities are expected to occur near home or within the same neighborhood. She has a strong nearest-option preference when choosing destinations, and tends to visit fixed local places repeatedly.

A.4 Prompt example

We provide two representative prompt examples: one for the mobility simulation model and another for action generation in the MDP-based prompt adjustment process.

Abridged mobility simulation prompt example

INPUT DATA

`Profile: {profile, today_date, home_location, work_location}.`
`Date: {today_date}.`

ROLE & TASK

You are a human mobility simulation agent. Given a person's profile, date, home location, work location, and candidate POIs, generate the person's daily mobility trajectory. The output should be a sequence of mobility events.

TASK REQUIREMENTS

- Generate events in chronological order.
- Each event should correspond to a physical visit to a location with a clear purpose.
- Do **not** generate passive or non-mobility activities, such as sleeping, resting at home, watching TV, or browsing the phone.
- Do **not** generate continuous GPS traces; each output item should be an event-level spatiotemporal point.
- ...

OUTPUT FORMAT

1. At 8:12 a.m., commute from home to the logistics warehouse for the morning work shift. Activity: Work. Location type: workplace. Location: [116.7, 38.0].

Abridged action generation prompt example

ROLE

You are an expert in human mobility modeling, urban science, statistical physics of mobility, and LLM-based synthetic population simulation...

BACKGROUND

The simulated data is created by feeding human personas (profiles) into an LLM, which generates mobility trajectories. The objective is to improve simulation realism by adjusting persona design or constraining behavioral rules...

INPUT DATA

ANALYSIS GOALS

- Population composition biases
- Missing or exaggerated activity-space archetypes
- Behavioral causes of distribution mismatch
- Ways to adjust personas or behavioral rules

TASK REQUIREMENTS

TASK 1 — Population-level Diagnosis

From a human mobility perspective, briefly explain what the observed differences suggest about overall activity space. Describe whether the simulated population tends to move within larger or smaller areas than real people, and identify which general mobility archetypes may be overrepresented or underrepresented...

TASK 2 — Population Group

Goal:Define a practical radius-based segmentation schema.
Requirements:

- Define 4–6 radius groups that partition individuals by activity-space size (small → large).
- Groups must be defined by radius magnitude.
- For each group provide: Group name, Radius range (clear thresholds or relative scale), One-sentence behavioral description and Target proportion (%) representing a realistic population distribution...

TASK 3 — Adjustment Strategy

For each group defined in Task 2, provide concrete adjustments aimed at improving simulation realism...
Examples include:

- Adding behavioral constraints (e.g., stronger home/-work anchors)
- Increasing routine or location stability...

OUTPUT FORMAT

(1) Population Diagnosis
(2) Population Groups Table
Columns: Group Name | Radius Range | Behavioral Description | Target proportion
(3) Adjustment Strategy (group-by-group suggestions)