AGENT-BASED HUMAN MOBILITY SIMULATION WITH PERSONALIZED COGNITIVE MAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models enable agents to simulate human mobility through sequential decisions about daily activities and destinations, without requiring task-specific training on large-scale trajectory data. However, existing methods often overlook how each individual's limited and evolving knowledge of the city constrains their mobility decisions. To address this limitation, we introduce CogMob, which equips each agent with a personalized, partial, and evolving cognitive map and incorporates it into mobility decision-making, where intent-aware retrieval activates relevant spatial knowledge and destination selection uses this knowledge to compare candidates in light of the agent's accumulated experience and current context. Experiments on two public datasets show that CogMob consistently improves mobility simulation across a wide range of metrics. We further assess its downstream utility through segregation analysis, observing patterns qualitatively consistent with prior findings. Our code is available at `https://anonymous.4open.science/r/CogMob-17B2`.

1 INTRODUCTION

Human mobility data are important for understanding travel demand and socioeconomic patterns, as well as informing transportation and urban planning Barbosa et al. (2018). However, real-world trajectory data are often difficult to obtain at scale, and their collection and sharing raise privacy concerns. These challenges motivate mobility simulation as a means of generating synthetic trajectories Luca et al. (2021). Among existing approaches, deep learning-based methods learn mobility patterns from observed trajectories to generate synthetic mobility sequences Deng et al. (2025); Feng et al. (2020). However, these methods typically require substantial real-world trajectory data for training, limiting their applicability when such data are scarce or inaccessible. Recent advances in LLM-based generative agents enable mobility simulation through explicit modeling of individual decisions JIAWEI et al. (2024); Du et al. (2025); Shao et al. (2026); Bhandari et al. (2024). These methods represent individuals as agents that use LLMs to reason about activities based on personal profiles and contextual information, generating mobility trajectories through sequential decisions.

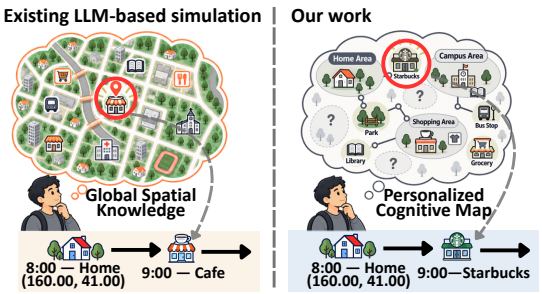


Figure 1: Core idea of CogMob

Despite this potential, existing LLM-based mobility simulation methods often make destination choices without explicitly accounting for each individual's accumulated knowledge of the city. For real individuals, however, knowledge of the urban environment constrains which options they consider and shapes their judgments of suitability and feasibility when planning activities and travel Kwan & Hong (1998); Hannes et al. (2008). This knowledge is shaped by personal experience: even within the same city, people with similar needs may know different places, be familiar with different areas, and perceive distance and accessibility differently. In cognitive science, such spatial knowledge is commonly conceptualized in terms of **cognitive maps**–internal representations of places and spatial relations that support navigation and spatial decision-making Epstein et al. (2017); Golledge (1999); Kuipers (1978). Such individualized spatial knowledge is not necessarily represented in existing

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

LLM-based mobility simulation methods. Some methods JIAWEI et al. (2024) use an LLM to generate destinations or select among candidate places, while others Du et al. (2025); Shao et al. (2024) combine LLM-generated activities with gravity-based assignment. These strategies can identify suitable places for an activity without explicitly tracking individuals' spatial knowledge or how it evolves through experience. What individuals know about the city and where they have been influence their mobility behavior, such as whether they explore new places and which familiar places they revisit. Our study (Section 2) shows that representative methods do not fully reproduce these behaviors: simulated individuals show less variation in how quickly they discover new places than real individuals, and their preference for returning to frequently visited places differs in strength from that observed in real data. This motivates us to equip each agent with a personalized cognitive map (Figure 1), making its knowledge of the city an explicit constraint on mobility decisions.

Building such cognitive map-guided mobility agents faces two challenges. (1) Cognitive map construction and evolution: A cognitive map is not given a priori, and our experiments show that simply using a collection of past visits as spatial memory leads to poorer simulation performance (Appendix D.4). Building such a map requires organizing an individual's partial spatial knowledge into a coherent structure and maintain this representation as new experiences change what is known and familiar. (2) Cognitive map-guided decision-making: A cognitive map guides mobility decisions by shaping which places an agent considers and how it chooses among place candidates. However, an agent may want to visit its usual café or simply want coffee without knowing where to go. This requires retrieving knowledge from the cognitive map relevant to its intention and combining that knowledge with its current needs and contextual constraints to guide destination choice.

We design CogMob, a cognitive map-guided framework for LLM-based human mobility simulation. To address the first challenge, CogMob represents each agent's partial spatial knowledge as a structured memory organized across three levels: places, routes, and regions. An incremental update mechanism incorporates newly visited and newly learned locations and refines existing knowledge, enabling the cognitive map to evolve with the agent's accumulated experience. To address the second challenge, CogMob incorporates the cognitive map into mobility decision-making through intent-guided retrieval and destination selection. Intent-guided retrieval dynamically activates relevant parts of the cognitive map according to the agent's varying spatial intents, producing a bounded set of candidate places. Destination selection then integrates the agent's accumulated spatial experience with its current activity context to select a destination.

In particular, our main contributions are as follows.

- Inspired by cognitive science research on human spatial decision-making, we explore the use of personalized cognitive maps to constrain agents' mobility behavior based on their partial and evolving knowledge of the urban environment.
- We design CogMob, a cognitive map-guided framework for LLM-based human mobility simulation. The framework maintains a personalized, partial, and evolving cognitive map for each agent and integrates it into mobility decision-making, where intent-aware retrieval activates relevant spatial knowledge and destination selection uses this knowledge to compare candidates in light of the agent's accumulated experience and current context.
- We conduct extensive experiments on two public datasets, where our method outperforms all training-free LLM-based baselines across all metrics, with average improvements of 49.8% on Beijing and 31.4% on NYC over the strongest baseline for each metric. Without real-world trajectory training, it also outperforms training-based baselines on most metrics.
- We demonstrate the potential of our synthetic mobility data for socioeconomic research through segregation analysis, which reveals patterns qualitatively consistent with prior empirical findings Moro et al. (2021).

2 MOTIVATION

Existing LLM-based mobility simulation methods often make destination choices without explicitly accounting for each individual's knowledge of the city. This knowledge may influence individuals' mobility decision and behavior, such as whether they explore new places and which familiar places they revisit. Without modeling these influences, plausible individual momentary choices may

still lead to unrealistic mobility patterns over time. We therefore examine exploration growth and preferential return Song et al. (2010a) as concrete examples.

We select two representative methods with different destination-selection strategies: gravity-based destination assignment Shao et al. (2026) and LLM-based selection JIAWEI et al. (2024) from candidate destinations, denoted as Gravity-assign and LLM-assign, respectively. Both methods are adapted to our simulation setting, with adaptation details provided in Appendix B.2. Using each method, we generate one-month trajectories for Beijing and compare them with real-world trajectories (data description in Section 4.1) from the same month. For each individual, we fit two curves: how the number of distinct visited locations increases over time, and how return probability varies with past visitation frequency. The exploration exponent μ describes how quickly the first curve grows, while the return exponent γ describes how strongly return probability increases with visitation frequency. We compare their distributions across individuals in real and simulated data. Definitions are in Appendix B.5.

As shown in Figure 2, both existing methods produce narrower exploration-exponent distributions than observed in real data, indicating insufficient heterogeneity in individual exploration behavior. The gravity-assign distribution is additionally shifted toward higher exponents, suggesting faster growth in the number of distinct visited locations. For preferential return, the return-exponent distribution shifts toward lower values for the LLM-assign method and higher values for the gravity-assign method, suggesting weaker and stronger preferences for frequently visited locations, respectively. These results show that, without accounting for what individuals know about the city, the simulation fail to fully reproduce individual differences in exploration and return patterns. These findings motivate us to incorporate personalized cognitive maps into simulation.

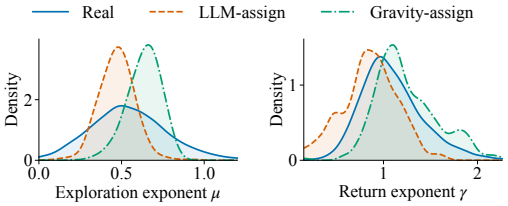


Figure 2: Distributions of exploration exponents (left) and preferential return exponents (right) from real and simulated trajectories in Beijing. Larger values in the x-axis indicate higher exploration growth and preferential return. The y-axis is the density.

3 METHODOLOGY

3.1 DEFINITION

Trajectory: A mobility trajectory refers to a sequence of discrete location visits, rather than a densely sampled GPS trace. Each visit records where an individual stays at a specific time. Formally, we define the trajectory of individual i on day d as $Y_d^i = \big((l_{d,1}^i, t_{d,1}^i), \ldots, (l_{d,T_{i,d}}^i, t_{d,T_{i,d}}^i)\big)$, where $l_{d,k}^i$ denotes the geographic coordinates of the k-th visit and $t_{d,k}^i$ denotes its continuous timestamp within the day. The trajectory length $T_{i,d}$ is not fixed and varies across individuals and days.

LLM-based mobility simulation. Given an individual's profile (e.g., demographics, occupation, and home/work locations) and daily context (e.g., date and day of the week), the task is to generate mobility trajectory Y_d^i using an LLM-based agent. We focus on a setting where the individual's real-world trajectories are unavailable, either as training supervision or as input during generation.

3.2 OVERVIEW

Figure 3 provides an overview of our framework. The left part illustrates the composition of a mobility agent and how it interacts with an urban environment to perform mobility simulation. We treat each individual as an LLM agent that maintains a persona, a need state, and a memory system consisting of episodic memory and a personalized cognitive map. Based on these components, the agent makes mobility decisions and executes them in the environment. After execution, the agent updates its internal state and memory based on environmental feedback about the activities performed, forming a continuous loop that generates the individual's mobility trajectory over time.

The right part details the agent's decision-making process. At each decision step, the agent follows three stages. First, *planning* determines the next activity and forms an initial idea of where to go based on the agent's profile, needs, memory, and current context. The agent may already have a particular destination in mind, such as a familiar gym, or only a broad preference, such as exercising

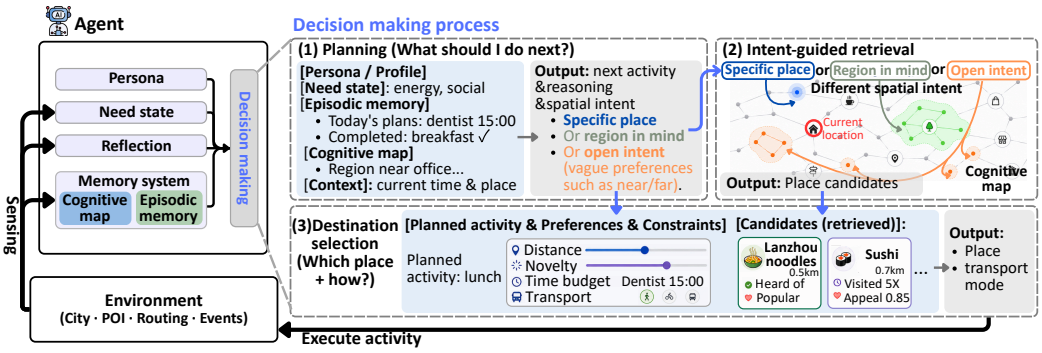


Figure 3: Framework of CogMob.

somewhere nearby. Second, *intent-guided retrieval* selectively activates relevant parts of the personalized cognitive map based on the agent's initial idea of where to go, yielding a bounded set of candidate places. Third, *destination selection* selects a destination from the retrieved candidates and determines a travel mode based on the planned activity, preferences, and constraints. The decision is then executed in the environment, whose feedback on the activities performed and locations visited updates the agent's internal state and memory for subsequent decisions.

3.3 Mobility Agent

Each agent comprises a persona, a need state, a reflection module, a memory system (including episodic memory and a personalized cognitive map), and a decision-making module, as illustrated in Figure 3. At initialization, the LLM expands a small set of given attributes (e.g., demographics and home/work locations) into a persona that includes behavioral preferences, daily routines, and personal interests, providing individualized context for subsequent mobility decisions (Details are provided in Appendix A.1). During simulation, the agent executes mobility decisions in the environment and uses environmental feedback to update its need state and memory. The following describes the need state, episodic memory, and periodic reflection. The personalized cognitive map and the decision-making process are detailed in Sections 3.4 and 3.5.

Need state: The agent tracks hunger, energy, and social demand, updating them after each travel or activity segment and applying overnight recovery between days. The current state is expressed in natural language and provided to the decision-making module to inform subsequent decisions. Detailed update rules are provided in Appendix A.2. **Episodic memory:** The episodic memory provides cross-day continuity by carrying past experience into future decisions. It contains four components. (1) Pending obligations are unfinished tasks that persist until completed or expired. (2) Upcoming plans record future commitments and become fixed anchors on the scheduled day. (3) Recent days maintains a rolling window of daily summaries, including mood and notable events. (4) Reflection insights capture higher-level observations from reflection. An example is provided in Appendix A.3. **Reflection:** Reflection periodically reviews the agent's accumulated experience to update its episodic memory and cognitive map. It takes two forms. First, a weekly review examines recent behavioral patterns, work intensity, mood, and frequently visited places to generate higher-level insights. These insights are stored in episodic memory and inform subsequent planning. Second, a personal place appeal review reassesses how appealing frequently visited places are to the individual. The updated appeal values are stored in the cognitive map and inform future destination choices. An example and a prompt are provided in Appendices A.3 and F.3.

3.4 Personalized Cognitive Map

Structure: The cognitive map forms the spatial component of the agent's memory. It is a structured and personalized representation of the city, containing knowledge acquired through initialization, direct visits, and indirect information. Inspired by human spatial knowledge at multiple scales Tversky (1993); Kuipers (1978; 2000), we organize it into three layers: places, routes, and regions.

(1) Place level (L_1). This level represents individual places in three groups. ***Visited places*** (L_{1a}) have been directly visited by the agent. Each place p stores coordinates, a category, visit count, and memory strength $s_p \in (0, 1]$. Additional attributes include personal place appeal updated through reflection and a record of the times of day when the agent typically visits the place. ***Known-of***

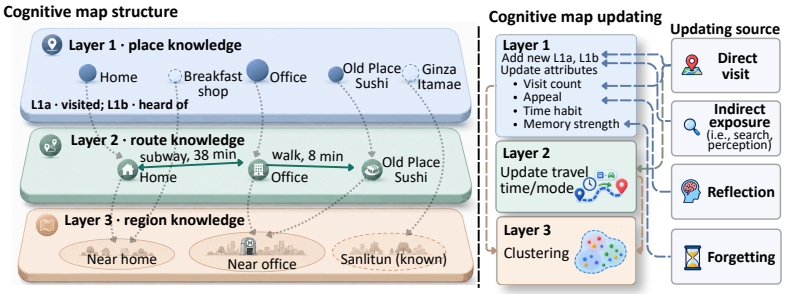


Figure 4: Cognitive Map Structure and Updating.

places (L_{1b}) are known but unvisited, acquired through search or indirect exposure. They store basic spatial and semantic information with a weaker familiarity signal. ***Anchors*** are home and work locations initialized from the profile and maintained separately as stable spatial references. (2) Route level (L_2). This level represents spatial connections learned through actual travel. Each route links an origin-destination pair and records information such as the travel mode and time. (3) Region level (L_3). This level represents broader areas in two forms. ***Visited regions*** summarize accumulated place-level experience and can be updated by periodically clustering visited places. ***Known-of regions*** are areas recognized through shared city knowledge, such as well-known commercial districts, without substantial direct experience. Both support region-level decisions, allowing the agent to choose a region before selecting a specific destination within it.

Update and forgetting: The cognitive map evolves as the agent interacts with the city. New information expands its knowledge, while unused knowledge gradually loses memory strength Epstein et al. (2017); Lever et al. (2002). (1) Updating. (i) Experience-driven updating. The cognitive map is updated through direct visits, indirect exposure, and reflection. First, direct visits update both place and route knowledge. Newly visited places are stored as visited places, while revisits increase visit counts and reinforce memory strength. The agent also records when each place is visited and stores the traveled origin–destination connection, travel mode, and travel time as route knowledge. Second, indirect exposure expands the set of known-of places. Locations discovered through candidate search or noticed during travel are stored as known-of places if they have not been visited. Once the agent visits one of these locations, it becomes a visited place. Third, reflection updates the agent's subjective evaluation of visited places. It reassesses personal place appeal based on accumulated experience, refining the preferences stored in the cognitive map without introducing new places. (ii) Region-level updating. As visited places accumulate, they are periodically clustered by location to form or update visited regions. (2) Forgetting. At the end of each day, the memory strength of each non-anchor place decays exponentially: $s_p \leftarrow s_p e^{-\lambda_p \Delta t}$, where λ_p is the decay rate and Δt is the elapsed time in days. Cognitive-map initialization is described in Appendix A.4. Detailed procedures and parameter settings for region-level updating and memory decay are in Appendix A.5.

3.5 Cognitive Map-Guided Mobility Decision-Making

At each decision point, the agent uses its personalized cognitive map to guide mobility decisions through three stages: *planning*, *intent-guided retrieval*, and *destination selection*.

Stage 1: Planning. People may think of an activity together with a specific destination, or decide what to do while still considering where to go. For example, someone may plan to have coffee at their usual café, or simply want a coffee somewhere quiet nearby without a particular place in mind. To capture this variation, the agent jointly generates the next activity and an initial location intent, conditioned on the agent's persona, current need state, episodic memory, and cognitive map. This intent may identify a particular place, indicate a broader region, or express only preferences concerning distance, novelty, and desired place attributes. An example prompt is in Appendix F.1.

Stage 2: Intent-guided retrieval. This stage models how an initial idea of where to go brings relevant spatial knowledge to mind. Guided by this intent, the agent selectively activates different parts of its personalized cognitive map to construct a bounded *consideration set* $\mathcal{C}$. (1) For a *specific-place intent*, such as visiting its usual café, the agent directly retrieves the place from its cognitive map. (2) For a *region-level intent*, such as looking for a café in Sanlitun, the agent activates that region and retrieves known places within it. (3) For an *open intent*, such as wanting coffee without a particular café in mind, retrieval is guided by the preferences available from planning. Two optional prefer-

ences affect different aspects of retrieval: distance defines the search radius, while novelty shapes the mix of candidate sources. A familiarity preference favors visited places, while a novelty preference favors known-of places and a few new POIs. Without an explicit novelty preference, retrieval combines visited places, known-of places, and new locations from the external POI database. Consulting the external POI database models how people use maps or apps to discover places beyond their existing knowledge. Details in Appendix A.6.

Stage 3: Destination selection. The agent selects a destination from $\mathcal{C}$ and a travel mode for the planned activity. Each candidate is described by its familiarity status, place attributes, distance, estimated travel time, visit history, and personal appeal when available. The agent selects a destination and travel mode by considering the candidate information alongside time budgets, available travel modes, and upcoming scheduled commitments. An example prompt is provided in Appendix F.2.

4 EVALUATION

We evaluate CogMob through baseline comparisons, ablations, cognitive-map augmentation, and segregation analysis. The appendix includes experiments on LLM backbones, social information interventions, and a baseline using visitation history, alongside individual footprint analyses and visual comparisons of real and simulated data.

4.1 EVALUATION SETUP

Dataset description: We use two public mobility datasets: a trajectory dataset from Beijing Shao et al. (2026) and a POI check-in dataset from New York City (NYC) Yang et al. (2015). These two datasets have been widely used to evaluate mobility modeling and prediction Guo et al. (2025). For the Beijing experiments, we simulate 1K users over one month. For NYC, we simulate 500 users over one month. Detailed dataset statistics and preprocessing procedures are in Appendix B.1.

Baselines: We compare our method against two categories of human mobility simulation baselines: **(1) LLM-based training-free methods**, including CoPB Shao et al. (2026), Urban-Mobility-LLM (UML) Bhandari et al. (2024), LLMob JIAWEI et al. (2024), CitySim Bougie & Watanabe (2025). **(2) and training-based methods**, including Geo-LLaMA Li et al. (2024), TimeGeo Jiang et al. (2016), MIRAGE Deng et al. (2025) and MoveSim Feng et al. (2020) that require model training or fine-tuning on real-world mobility data. The baseline details are in Appendix B.2.

Metric: We evaluate the simulation from three perspectives using metrics widely adopted in prior mobility simulation studies Cornacchia & Pappalardo (2021); Yuan et al. (2026). Many of the mobility patterns measured by these metrics have been extensively studied and repeatedly observed in real-world data Gonzalez et al. (2008). (1) First, we examine whether simulated individuals reproduce the fundamental mobility characteristics observed in real data, including **stay duration**, **jump distance**, **radius of gyration**, and **visitation entropy** Gonzalez et al. (2008); Song et al. (2010b). (2) Second, we assess mobility dynamics, examining how individuals explore new places, revisit familiar ones, and develop their set of activity locations over time. We measure these patterns using **exploration**, **return**, **visitation frequency**, and **capacity growth** Song et al. (2010a); Alessandretti et al. (2018). (3) Third, we assess urban spatial patterns by examining whether the simulated population reproduces the spatial distribution of visits (**LocFreq**) and origin–destination connections (**OdSim**) Shao et al. (2026) observed in real-world trajectories. For all metrics, we use Jensen–Shannon divergence (JSD) to quantify discrepancies between simulated and real-world data. Detailed definitions are provided in the Appendix B.5.

4.2 OVERALL PERFORMANCE

Table 1 compares the performance of CogMob and LLM-based training-free baselines on the Beijing and NYC datasets. Given space limitations, results including standard deviations across runs are provided in Appendix C.1. CogMob achieves the lowest JSD across all ten metrics on both datasets when compared with those produced from real data. Compared with the best-performing baseline for each metric, CogMob reduces JSD by 19.06%–81.43% on Beijing and 3.34%–74.70% on NYC. The improvements in mobility characteristics and dynamics indicate closer reproduction of movement and stay patterns, as well as exploration and revisitation behavior. Lower errors in urban spatial patterns further indicate better reproduction of the citywide distribution of visits and OD flows.

Table 1: Comparison with **training-free** baselines on Beijing and NYC. Bold scores indicate the best values. More detailed results including standard deviations are provided in Appendix C.1.

Dataset	Method	Mobility Characteristics				Mobility Dynamics				Urban Spatial Patterns	
		Radius	Jump	Duration	Entropy	Explore	Return	Visit-Freq	Capacity	LocFreq	OdSim
BJ	UML	0.0403	0.0716	0.2239	0.4181	0.2705	0.2929	0.5343	0.0031	0.1996	0.3636
	LLMob	0.0585	0.1486	0.2582	0.3427	0.1840	0.1231	0.1951	0.0046	0.2414	0.4675
	CoPB	0.1631	0.0574	0.2036	0.2834	0.1200	0.0528	0.1259	0.0018	0.3057	0.4404
	CitySim	0.0518	0.0709	0.3966	0.3467	0.1504	0.0568	0.1424	0.0013	0.2802	0.4562
	CogMob	**0.0214**	**0.0292**	**0.0378**	**0.1331**	**0.0271**	**0.0295**	**0.0911**	**0.0003**	**0.1544**	**0.2943**
NYC	UML	0.3988	0.1251	0.3712	0.5849	0.2274	0.1104	0.2473	0.0031	0.1220	0.3950
	LLMob	0.3222	0.2567	0.3478	0.5697	0.3770	0.2420	0.4773	0.0165	0.1107	0.4470
	CoPB	0.2727	0.0894	0.3239	0.1680	0.0320	0.2201	0.3209	0.0026	0.1212	0.3939
	CitySim	0.4186	0.1234	0.3462	0.1445	0.1248	0.2227	0.2346	0.0083	0.1265	0.3924
	CogMob	**0.0690**	**0.0469**	**0.2381**	**0.1319**	**0.0295**	**0.0396**	**0.0919**	**0.0023**	**0.1070**	**0.3582**

4.3 Comparison with Training-Based Methods

We also compare CogMob with baselines trained on real mobility data, as shown in Table 2. Detailed experimental settings and results are provided in Appendix C.2. From the Table, we make the following observations. (1) CogMob achieves the lowest JSD on most metrics across both datasets, particularly in mobility dynamics, despite the baselines' access to real mobility data. These gains may partly reflect its use of day-specific context and accumulated experience to maintain behavioral continuity across days. (2) Beijing trajectories contain routine local mobility with occasional long-distance intercity travel. Such long-tailed travel-distance distributions have been documented in mobility research Brockmann et al. (2006). The training-based baselines appear sensitive to these rare long-distance trips, overestimating typical travel distances and individuals' activity ranges. In contrast, CogMob more accurately reproduces both typical mobility and long-distance tails without real-world trajectory training. A detailed analysis is provided in Appendix C.2. Accurately capturing both routine and long-distance mobility is important to transportation planning Chen et al. (2023), emissions estimation Wadud et al. (2024), and epidemic simulation Lai et al. (2020). (3) The baselines remain competitive on urban spatial patterns. Training on real trajectories may help them learn which locations are frequently visited and which locations people commonly travel between.

Table 2: Comparison with **training-based** baselines on Beijing and NYC. Bold scores indicate the best values. More detailed results including standard deviations are provided in Appendix C.2

Dataset	Method	Mobility Characteristics				Mobility Dynamics				Urban Spatial Patterns	
		Radius	Jump	Duration	Entropy	Explore	Return	Visit-Freq	Capacity	LocFreq	OdSim
BJ	Geo-LLaMA	0.3115	0.1538	0.1264	0.1587	0.0668	0.0813	0.1023	0.0022	0.2210	0.4359
	Movesim	0.1482	0.2317	0.1595	**0.0783**	0.1613	0.1043	0.2738	0.0147	0.3924	0.7310
	MIRAGE	0.3390	0.4222	0.1032	0.3419	0.0809	0.0804	0.1878	0.0019	0.1333	0.6486
	TimeGeo	0.0808	0.1897	0.0782	0.2995	0.0919	0.0378	0.1715	0.0027	**0.1049**	0.4476
	CogMob	**0.0214**	**0.0292**	**0.0378**	0.1331	**0.0271**	**0.0295**	**0.0911**	**0.0003**	0.1544	**0.2943**
NYC	Geo-LLaMA	0.1599	0.0856	0.4279	0.3654	0.1659	0.2193	0.1874	0.0032	0.1917	0.6295
	Movesim	0.2524	0.0711	0.4107	0.3869	0.2238	0.2015	0.3260	0.0101	0.2115	0.6016
	MIRAGE	0.1785	0.0986	0.2439	0.1425	0.1040	0.0796	0.1012	0.0040	0.0325	0.3651
	TimeGeo	**0.0366**	**0.0116**	0.3313	0.2851	0.0697	0.2041	0.1299	0.0281	**0.0239**	**0.3208**
	CogMob	0.0690	0.0469	**0.2381**	**0.1319**	**0.0295**	**0.0396**	**0.0919**	**0.0023**	0.1070	0.3582

4.4 Ablation Study

We first examine the role of the cognitive map through two variants (w/o CM-A and w/o CM-B): Both variants remove the cognitive map but retain episodic memory, allowing past activities and experiences to inform subsequent decisions. w/o CM-A lets the LLM directly generate destinations, while w/o CM-B retrieves real-world POIs near the current location. As shown in Table 3, providing real POIs in w/o CM-B reduces many of the errors observed in w/o CM-A. However, the full model generally achieves lower JSD than both variants on mobility characteristics and mobility dynamics. These results suggest that personalized cognitive maps provide benefits beyond episodic summaries and real POI candidates by explicitly modeling individuals' spatial knowledge.

Table 3: Ablation Study on Beijing datasets. Bold scores indicate the best values.

Method	Mobility Characteristics				Mobility Dynamics				Urban Spatial Patterns	
	Radius	Jump	Duration	Entropy	Explore	Return	Visit-Freq	Capacity	LocFreq	OdSim
CogMob w/o CM-A	0.0322	0.1278	**0.0327**	0.1983	0.2523	0.6114	0.7636	0.0027	0.4377	0.6210
CogMob w/o CM-B	0.0264	0.0372	0.0462	0.1847	0.1001	0.1213	0.1581	0.0004	0.1527	**0.2935**
CogMob w/o DR	0.0291	0.0384	0.0400	0.1677	0.0326	0.0302	0.1050	0.0004	**0.1506**	0.3006
CogMob	**0.0214**	**0.0292**	0.0378	**0.1331**	**0.0271**	**0.0295**	**0.0911**	**0.0003**	0.1544	0.2943

We next examine the role of dynamic retrieval. In w/o DR, candidates are retrieved from the same nearby pool regardless of the agent's intent. Removing intent-dependent retrieval increases JSD

across all mobility characteristics and mobility dynamics metrics. This suggests that adapting retrieval to the current intent helps agents use their cognitive maps more effectively when making mobility decisions. For urban spatial pattern metrics, the full model performs comparably to w/o CM-B and w/o DR, with only small differences in LocFreq and OdSim JSD. Ablation variants and results with standard deviations across runs are detailed in Appendices B.3 and C.3, respectively.

4.5 Cognitive Maps as a Plug-and-Play Augmentation

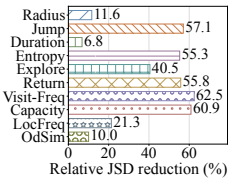


Figure 5: LLMob's JSD reduction with CogMob's map.

Beyond evaluating CogMob as a standalone mobility simulator, we investigate whether its cognitive map can augment an existing simulation framework as a plug-and-play component. We select LLMob as the testbed because it relies on real check-in histories to provide candidate locations and short-term mobility context. Since such histories are unavailable in our settings, LLMob offers a natural setting for evaluating whether CogMob's cognitive map can provide an effective alternative. We compare LLMob with LLMob augmented by CogMob's cognitive map (LLMob w CM). In LLMob w CM, LLMob selects candidate locations from each agent's CogMob cognitive map, including places the agent has visited and places it knows about. Figure 5 shows that cognitive-map augmentation (LLMob w CM) reduces LLMob's JSD across all metrics. These results suggest that incorporating personalized spatial knowledge helps agents better reproduce real-world mobility patterns, while demonstrating that CogMob's cognitive map can improve an existing simulator.

4.6 Application of Simulated Mobility Data

This section demonstrates the utility of simulated data for downstream applications, focusing on socioeconomic segregation analysis. Socioeconomic segregation refers to the separation of people from different socioeconomic groups Athey et al. (2020); Nilforoshan et al. (2023). In daily activities, individual experienced segregation describes individuals' uneven exposure to different socioeconomic groups. Prior studies link this segregation to mobility behavior, particularly place and social exploration Moro et al. (2021). Motivated by these works, we examine whether similar segregation patterns emerge in simulated data. We present the socioeconomic segregation for NYC here and report the remaining results in Appendix C.4.

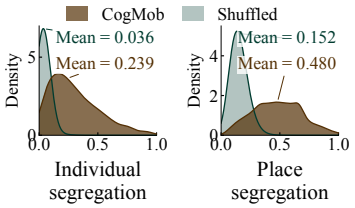


Figure 6: Individual and place segregation NYC simulations vs. shuffled.

Our first question is: *Do simulated mobility data exhibit non-random socioeconomic segregation?* We measure (1) place segregation, the concentration of visitors' time across income groups at a location, and (2) individual segregation, the concentration of an individual's exposure across income groups, weighted by time spent at visited locations. Higher values indicate greater socioeconomic segregation. To examine whether simulated data exhibits segregation and whether it is stronger than under random destination choices, we follow the prior work and shuffle visit destinations within the simulated NYC trajectories, keeping each person's visit counts and durations unchanged. We then compare both segregation measures between the original (i.e., CogMob) and randomized trajectories (i.e., Shuffled). Figure 6 compares the distributions of individual experienced segregation (left) and place segregation (right) between simulated and randomized trajectories. Segregation varies widely across individuals and places, with both measures higher overall in the simulated data. These patterns are consistent with the non-random socioeconomic segregation observed in prior work Moro et al. (2021). We then assess how closely these simulated segregation distributions resemble those observed in real mobility data. Comparing them with the pooled public data from 11 cities in prior work Moro et al. (2021), we obtain JSD values of 0.0099 for individual experienced segregation and 0.0138 for place segregation. These low values indicate similar distributions.

Our second question is: *Can the association between exploration behavior and experienced segregation also be observed in simulated mobility data?* Following the prior work, we measure (1) place exploration, the number of distinct places visited relative to the total number of visits, and (2) social exploration, the fraction of visited places where the individual's own income group is not the majority. We then examine how these two types of exploration are associated with experienced segregation in the simulated data. Figure 7 shows mean individual experienced segregation (color) across social exploration (horizontal axis) and place exploration (vertical axis), for simulated NYC data (left) and data from prior work Moro et al. (2021) (right). In

the simulated data, higher levels of both exploration measures are generally associated with lower mean individual experienced segregation, consistent with the pattern observed in the prior work. For a quantitative comparison with the data from 11 cities in prior work, we obtain a three-dimensional JSD of 0.1192 for the joint distribution of place exploration, social exploration, and individual experienced segregation. Together with the visual trend, this result supports similar exploration–segregation patterns. Our method also achieves lower JSD than the best LLM-based training-free baseline (i.e., CoPB), with detailed comparisons provided in Appendix C.4.

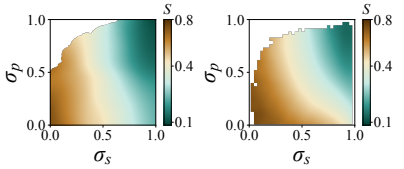


Figure 7: Exploration and individual segregation in NYC simulation (left) and prior work (right) Moro et al. (2021).

5 Related Work

LLM-based Mobility Simulation: Recent studies use LLM-based agents to generate mobility trajectories through sequential decisions JIAWEI et al. (2024); Du et al. (2025); Piao et al. (2025); Mou et al. (2024); Liu et al. (2024); Ju et al. (2025); Shao et al. (2026); Bhandari et al. (2024); Li et al. (2024). For example, CoPB Shao et al. (2026) combines intention and activity planning with gravity-based destination assignment, while LLMob JIAWEI et al. (2024) uses self-consistency and retrieval strategies to guide personal mobility generation. However, these approaches often overlook the cognitive constraints imposed by individuals' partial and evolving knowledge of the city. Our framework explicitly represents this knowledge through personalized cognitive maps and incorporates these constraints into agents' mobility decision-making.

Agent Memory: Existing agent-memory systems retain and reuse past interactions, experiences, knowledge, or skills for long-horizon reasoning and adaptation Zhong et al. (2024); Rasmussen et al. (2025); Huang et al. (2026); Park et al. (2023). Most focus on dialogue, task execution, or tool use without explicitly modeling an agent's spatial knowledge. Some address spatial reasoning, but at different environmental scales and with different objectives. In bounded indoor environments, agents can build relatively comprehensive grid maps or spatial graphs for localization and navigation Wang et al. (2025a; 2023b); Gupta et al. (2026); Wang et al. (2025b); Singh et al. (2023). In larger, open-world environments, agents such as Voyager focus on task completion and skill acquisition, organizing memory around experiences and reusable skills rather than explicit spatial knowledge Wang et al. (2023a; 2024). Our setting poses a different challenge: it requires modeling how individuals' partial knowledge of the city constrains mobility decisions. Ignoring this knowledge, or assuming it is complete or identical across agents, can lead to unrealistic mobility patterns. We therefore represent each agent's spatial knowledge through a personalized cognitive map.

Cognitive Map: Research in spatial cognition and neuroscience shows that humans develop internal spatial representations through experience to guide navigation and spatial decisions Golledge (1999); Pappalardo et al. (2015). These representations, known as cognitive maps Burgess et al. (2002); Behrens et al. (2018), suggest that mobility decisions depend not only on the physical environment but also on an individual's internal spatial knowledge, motivating us to equip each agent with a personalized, evolving cognitive map. Cognitive science further informs this structure: human spatial knowledge is often partial and uneven Tversky (1993). It can be organized hierarchically around places and regions, ranging from local topological relations to integrated geometric representations Hirtle & Jonides (1985); Kuipers (1978; 2000). Cognitive-map-inspired representations have been studied for navigation and planning in embodied AI and robotics Banino et al. (2018); Gupta et al. (2017); Chaplot et al. (2020). Unlike these works, which mainly represent explored environments to support navigation, our setting requires modeling each individual's partial and personalized urban knowledge to guide mobility decisions at city scale.

6 Conclusion

We introduced CogMob, a cognitive map-guided framework for LLM-based human mobility simulation. It maintains a personalized, partial, and evolving cognitive map for each agent and uses intent-guided retrieval to incorporate relevant spatial knowledge into mobility decision-making. Experiments on Beijing and NYC show that CogMob consistently improves mobility simulation across a wide range of metrics. We further assess its downstream utility through segregation analysis, which exhibit segregation patterns qualitatively consistent with prior findings.

AI USE STATEMENT

As our study focuses on LLM-based human mobility simulation, we used LLMs as agent backbones in CogMob and the evaluated LLM-based baselines to generate synthetic mobility data, as described in the methodology and experimental settings.

For research and evaluation, generative AI tools assisted with methodological discussion, development and revision of research code, and interpretation of results. For manuscript preparation, they assisted with translation, drafting and editing, and figure revision. The authors take responsibility for the final manuscript, code, and reported results.

REFERENCES

Laura Alessandretti, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli. Evidence for a conserved quantity in human mobility. *Nature human behaviour*, 2(7):485–491, 2018.

Amap. Amap Web Service API: POI Search. `https://developer.amap.com`. Accessed: 2026-09.

Susan Athey, Billy A Ferguson, Matthew Gentzkow, and Tobias Schmidt. Experienced segregation. Technical report, National Bureau of Economic Research, 2020.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

Timothy EJ Behrens, Timothy H Muller, James CR Whittington, Shirley Mark, Alon B Baram, Kimberly L Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018.

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

Nicolas Bougie and Narimawa Watanabe. 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*, pp. 215–229, 2025.

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

Neil Burgess, Eleanor A Maguire, and John O'Keefe. The human hippocampus and spatial and episodic memory. *Neuron*, 35(4):625–641, 2002.

Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological slam for visual navigation. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12872–12881. IEEE, 2020.

Yong Chen, Haoge Xu, Xiqun Michael Chen, and Ziyou Gao. A multi-scale unified model of human mobility in urban agglomerations. *Patterns*, 4(11), 2023.

Giuliano Cornacchia and Luca Pappalardo. Sts-epr: Modelling individual mobility considering the spatial, temporal, and social dimensions together. *Procedia Computer Science*, 184:258–265, 2021.

Bangchao Deng, Xin Jing, Tianyue Yang, Bingqing Qu, Dingqi Yang, and Philippe Cudre-Mauroux. Revisiting synthetic human trajectories: Imitative generation and benchmarks beyond datasaurus. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 201–212, 2025.

Yuwei Du, Jie Feng, Jian Yuan, and Yong Li. Cams: A citygpt-powered agentic framework for urban human mobility simulation. *arXiv preprint arXiv:2506.13599*, 2025.

Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience*, 20(11):1504–1513, 2017.

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

Reginald G Golledge. *Wayfinding behavior: Cognitive mapping and other spatial processes*. JHU press, 1999.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Baoshen Guo, Zhiqing Hong, Lidan Cao, Donghang Li, Junyi Li, Can Rong, Alok Prakash, Shenhao Wang, and Jinhua Zhao. Language models meet urban mobility: A data-centric review. 2025.

Gunshi Gupta, Karmesh Yadav, Zsolt Kira, Yarin Gal, and Rahaf Aljundi. Memo: Training memory-efficient embodied agents with reinforcement learning. *Advances in Neural Information Processing Systems*, 38:104392–104414, 2026.

Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2616–2625, 2017.

Els Hannes, Davy Janssens, and Geert Wets. Destination choice in daily activity travel: mental map's repertoire. *Transportation research record*, 2054(1):20–27, 2008.

Stephen C Hirtle and John Jonides. Evidence of hierarchies in cognitive maps. *Memory & cognition*, 13(3):208–217, 1985.

Wei-Chieh Huang, Weizhi Zhang, Yueqing Liang, Yuanchen Bei, Yankai Chen, Tao Feng, xinyu Pan, Zhen Tan, Yu Wang, Tianxin Wei, Shanglin Wu, Ruiyao Xu, Liangwei Yang, Rui Yang, Wooseong Yang, Chin-Yuan Yeh, Hanrong Zhang, Haozhen Zhang, Siqi Zhu, Henry Peng Zou, Wanjia Zhao, Song Wang, Wujiang Xu, Zixuan Ke, Zheng Hui, Dawei Li, Yaozu Wu, Langzhou He, Chen Wang, Xiongxiao Xu, Baixiang Huang, Juntao Tan, Shelby Heinecke, Huan Wang, Caiming Xiong, Ahmed Metwally, Jun Yan, Chen-Yu Lee, Hanqing Zeng, Yinglong Xia, Xiaokai Wei, Ali Payani, Yu Wang, Haitong Ma, Wenya Wang, Chenguang Wang, Yu Zhang, Xin Eric Wang, Yongfeng Zhang, Jiaxuan You, Hanghang Tong, Xiao Luo, Xue Liu, Yizhou Sun, Wei Wang, Julian McAuley, James Zou, Jiawei Han, Philip S. Yu, and Kai Shu. A survey of agent memory in the second half: Towards self-evolving and long-horizon agents. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=XycbogUAeJ. Survey Certification.

Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C González. The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences*, 113(37):E5370–E5378, 2016.

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

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

Benjamin Kuipers. Modeling spatial knowledge. *Cognitive science*, 2(2):129–153, 1978.

Benjamin Kuipers. The spatial semantic hierarchy. *Artificial intelligence*, 119(1-2):191–233, 2000.

Mei-Po Kwan and Xiao-Dong Hong. Network-based constraints-oriented choice set formation using gis. *Geographical Systems*, 5:139–162, 1998.

Shengjie Lai, Nick W Ruktanonchai, Liangcai Zhou, Olivia Prosper, Wei Luo, Jessica R Floyd, Amy Wesolowski, Mauricio Santillana, Chi Zhang, Xiangjun Du, et al. Effect of non-pharmaceutical interventions to contain covid-19 in china. *nature*, 585(7825):410–413, 2020.

Colin Lever, Tom Wills, Francesca Cacucci, Neil Burgess, and John O'Keefe. Long-term plasticity in hippocampal place-cell representation of environmental geometry. *Nature*, 416(6876):90–94, 2002.

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

Yifan Liu, Xishun Liao, Haoxuan Ma, Brian Yueshuai He, Chris Stanford, and Jiaqi Ma. Human mobility modeling with household coordination activities under limited information via retrieval-augmented llms. *arXiv preprint arXiv:2409.17495*, 2024.

Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)*, 55(1):1–44, 2021.

Esteban Moro, Dan Calacci, Xiaowen Dong, and Alex Pentland. Mobility patterns are associated with experienced income segregation in large us cities. *Nature communications*, 12(1):4633, 2021.

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

Hamed Nilforoshan, Wenli Looi, Emma Pierson, Blanca Villanueva, Nic Fishman, Yiling Chen, John Sholar, Beth Redbird, David Grusky, and Jure Leskovec. Human mobility networks reveal increased segregation in large cities. *Nature*, 624(7992):586–592, 2023.

OpenStreetMap. OpenStreetMap. `https://www.openstreetmap.org`. Accessed: 2026-09.

Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature communications*, 6(1):8166, 2015.

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

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

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

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

Chenyang Shao, Bingbing Fan, Jingtao Ding, Yuan Yuan, Meng Wang, and Fengli Xu. Generating realistic human mobility data with hybrid large language model agent. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2758–2769, 2026.

Kunal Pratap Singh, Jordi Salvador, Luca Weihs, and Aniruddha Kembhavi. Scene graph contrastive learning for embodied navigation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10850–10860. IEEE, 2023.

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

Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010b.

Barbara Tversky. Cognitive maps, cognitive collages, and spatial mental models. In *European conference on spatial information theory*, pp. 14–24. Springer, 1993.

U.S. Census Bureau. 2009–2013 American Community Survey 5-year estimates: Table B19013, median household income, 2014. URL `https://data.census.gov`.

Zia Wadud, Muhammad Adeel, and Jillian Anable. Understanding the large role of long-distance travel in carbon emissions from passenger travel. *Nature Energy*, 9(9):1129–1138, 2024.

Bo Wang, Jiehong Lin, Chenzhi Liu, Xinting Hu, Yifei Yu, Tianjia Liu, Zhongrui Wang, and Xiaojuan Qi. Mg-nav: Dual-scale visual navigation via sparse spatial memory. *arXiv preprint arXiv:2511.22609*, 2025a.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15579–15590. IEEE, 2023b.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1894–1907, 2024.

Zixuan Wang, Bo Yu, Junzhe Zhao, Wenhao Sun, Sai Hou, Shuai Liang, Xing Hu, Yinhe Han, and Yiming Gan. Karma: Augmenting embodied ai agents with long-and-short term memory systems. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8. IEEE, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Dingqi Yang, Daqing Zhang, Vincent. W. Zheng, and Zhiyong Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2015. ISSN 2168-2216.

Yuan Yuan, Yuheng Zhang, Jingtao Ding, and Yong Li. Worldmove, a global open data for human mobility. *Scientific Data*, 13(1):549, 2026.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

A Details of the mobility agent

A.1 Persona Construction

Each agent's persona is constructed once at initialization from a minimal profile containing age, gender, occupation, education, income, and home/work coordinates.

Behavioral tendencies. We assign behavioral tendencies to capture individual differences in everyday behavior, guiding persona generation, mobility decisions, and the development of personalized cognitive maps. We characterize each agent using five behavioral tendencies: preference for cooking at home, curiosity about new experiences, interest in socializing, preference for regular routines, and general energy level. We predefine seven persona archetypes, each specifying a sampling range for each of the five tendencies. For example, a "social explorer" is more inclined to explore and socialize, but less inclined to follow a strict routine. For each individual, we randomly select one of the seven archetypes with equal probability. We then sample each behavioral score independently and from the range specified by that archetype. We subsequently apply age-dependent adjustments to exploration and energy. The resulting scores guide the LLM in generating consistent schedules, habits, and preferences at initialization and remain fixed throughout simulation.

We explain below why behavioral tendencies are assigned through programmatic sampling rather than generated directly by the LLM. In our preliminary study, we tested two ways of using the LLM to assign behavioral tendencies. When asked to generate the five scores directly, it assigned moderate values near 0.5 across all dimensions to 88.7% of individuals. This concentration persisted even with explicit instructions to avoid defaulting to 0.5. When instead asked to choose among seven predefined archetypes, the LLM selected "disciplined and efficient" for 44.7% of individuals and a homebody archetype for 26.0%, while rarely selecting social-explorer or foodie-explorer profiles. The first approach produces mostly moderate behavioral scores, while the second favors only a few archetypes. Without external evidence to calibrate the archetype proportions, we adopt uniform sampling to provide broader coverage of the predefined profiles. This is an initialization choice, rather than an estimate of real population proportions. In future work, we plan to incorporate external knowledge to inform and refine these sampling probabilities.

LLM-generated schedules and personal attributes. Conditioned on the basic profile and sampled behavioral tendencies, the LLM generates the following five groups of attributes:

- *Work schedule:* working days, office attendance days, work start/end times, lunch-break availability, work arrangement (office, home, or hybrid), weekday living location, overtime habits, and schedule notes. The prompt specifies occupation-dependent scheduling patterns, such as extended hours for finance workers, early starts for teachers, weekend work for service workers, and night shifts for bar staff.
- *Daily rhythm:* weekday and weekend wake times, a typical bedtime, and a short description of the individual's daily rhythm. These attributes guide the timing of waking and sleeping during simulation.
- *Meal habits:* separate descriptions for breakfast, lunch, and dinner, including whether the individual typically prepares food at home or eats out and what they prefer to eat.
- *Monthly travel plans:* whether the individual plans to leave the city and, for each planned trip, its timing, duration in days, purpose (leisure, business, or family visit), and destination. Destinations are specified as real cities, and the plans are subsequently converted into multi-day travel events.
- *Personal details:* weekend style, one to four specific interests, two to four habitual behaviors or quirks, and a 50–100-word personality description.

Illustrative example. The box below presents an abridged persona for an office worker in NYC. Conditioned on the basic profile and behavioral tendencies, the LLM generates concrete schedules, habits, travel plans, and personal details.

Persona example

An abridged persona presented in natural language. The LLM generates the schedules, habits, travel plans, and personal details conditioned on these tendencies and the individual's basic profile.

[Resident background]

An office worker in his early thirties, living in Washington Heights and working as an office assistant at an accounting firm on the Upper West Side.

[Sampled behavioral tendencies]

Prototype: free-spirited and loose.

Cooking: 0.241 · exploration: 0.625 · social: 0.580.
Routine strictness: 0.255 · energy: 0.359.

[Work schedule]

Works in the office Monday–Friday, 09:00–17:00, with a lunch break. Lives at home throughout the week and commutes approximately 20 minutes to work. Usually leaves on time unless the day is particularly busy, prioritizing time for social plans.

[Daily rhythm]

Wake time: 08:00 on weekdays; 10:00 on weekends.

Typical bedtime: 01:00.

Maintains a loose routine outside work, stays up late, and sleeps in on weekends.

[Meal habits]

Breakfast: usually grabs coffee and a croissant on the way to work.

Lunch: packs a sandwich or buys a quick meal from a nearby deli.

Dinner: orders takeout or picks up a slice from the local pizzeria.

[Monthly travel plan]

A two-day leisure trip to Philadelphia in the middle of the month to spend time with old friends from his hometown.

[Interests and personal habits]

Enjoys music, urban culture, and food. Weekends include relaxing, catching up with friends, and attending small concerts in Brooklyn.

Wears mismatched socks as a personal style, prefers working without a tie, and often leaves his phone charger on his desk.

[Personality summary]

A laid-back office worker with a free-spirited approach to life. He enjoys New York's music scene and urban culture, making time for spontaneous plans and late-night activities with friends.

A.2 Need State Updates

The need-state update mechanism consists of three parts: changes due to elapsed time, the effects of activities, and overnight recovery. First, elapsed time increases hunger and reduces energy. Energy consumption depends on activity type and duration, with travel and work consuming more energy than other activities, and is modulated by the agent's persona-derived stamina. Social demand increases during periods without social interaction. Second, activities produce additional changes to the need state. Meals reduce hunger, with smaller reductions for coffee or snacks; rest restores energy according to its duration; and social activities reduce social demand. Third, overnight recovery prepares the need state for the next simulated day. Energy is restored according to sleep duration and stamina, hunger is reset to a morning level, and the record of days without social interaction is updated. The persona's energy tendency modulates energy consumption and overnight recovery; all other update parameters are shared across agents.

Before each planning step, the current need state is translated into natural-language descriptions, such as "very hungry" or "somewhat tired," and included in the prompt to inform subsequent activity and destination choices.

A.3 Episodic Memory Example

The following example is drawn from the stored memory of NYC agent during the June 2012 simulation. Entries are abridged for readability. The example illustrates how unfinished tasks, future commitments, daily summaries, and reflection insights are retained across days.

Episodic memory example: NYC agent

[Pending obligations]

Pick up dry cleaning from Queens.
Activity: errand · Location: Long Island City.
Created: June 20 · Deadline: June 27 · Urgency: high.

Check out a live music venue in Williamsburg.
Activity: entertainment · Location: Williamsburg, Brooklyn.
Created: June 22 · Deadline: June 29 · Urgency: high.

[Upcoming plans]

Eat with friends at Mexico 2000.
Scheduled: July 1, 21:36 · Activity: social.
Location: Williamsburg · Created: June 30.

[Recent-day summaries]

One entry from the rolling 14-day window is shown.

Date: June 30, 2012.

Mood: excited from social outings but tired from a long day and travel.

Notable events: brunch in Williamsburg; visits to Java Street Garden and the National Jazz Museum in Harlem; a late-night social event at Mexico 2000.

Unfulfilled intention: pick up dry cleaning from Queens.

Daily activity summary: 14 activities, including four coffee activities, three leisure activities, and one social activity; 120 minutes of travel and no work.

Other daily summaries omitted.

[Reflection insights]

Pattern: "Visited Seabreeze Café eight times in the past week; maybe want to vary coffee spots."
Evidence: eight visits to Seabreeze Café.
Created: June 24 · Importance: 6.5.

Travel burden: "Spent 1,417 minutes on travel in the past week; high travel burden."
Evidence: 1,417 minutes of travel · Importance: 5.5.

Emerging need: "Mood fatigued seven days straight; need to schedule a real rest day next week."
Evidence: seven days of fatigue.
Created: June 17 · Importance: 5.0.

⋮

A.4 Cognitive-Map Initialization (Cold Start)

Before simulation begins, each agent receives a small cognitive map representing its initial knowledge of the city. We construct this map from the agent's profile, and public city information, without using the individual's real-world trajectories. The three levels are initialized as follows.

Place level. We initialize three types of places. (1) *Anchors:* For Beijing, home and work locations come from the provided profiles. For NYC, home locations come from our synthetic profiles, while

workplaces are assigned based on each individual's industry. The agent's work schedule determines whether it works at home or at the workplace on a given day.

(2) *Visited places:* We assume that each agent has visited a few everyday places near home and work before simulation begins. To initialize these familiar places, we collect nearby POIs and group them by category. Around home, we select up to three dining places, two daily shopping places, and one each for transport, leisure, and services. Around work, we select up to two dining places and one each for daily shopping and transport. To determine which places to include in each category, we assign higher scores to closer POIs, with small bonuses for available ratings and the agent's lifestyle tags. We then sample from the highest-scoring candidates, giving those with higher scores a greater chance of selection.

(3) *Known-of places:* We add well-known landmarks and representative places from city districts to represent locations the agent has heard of but has not visited. For example, a resident may know about the Empire State Building in NYC. We obtain these places from public webpages, maps, and review websites. The lists contain 41 landmarks in Beijing and 25 in NYC. Agents in each city share the same candidate list, but each receives a randomly sampled subset, allowing their initial knowledge of these places to vary.

Route level. At initialization, we provide travel information for trips between home and work, when applicable, and between these anchors and the nearby places initialized as visited. Routes in both directions are stored separately. Each route records how the agent would travel and how long the trip would take. We choose a travel mode from the agent's available transportation options and estimate the travel time using the routing service.

Region level. Known-of regions represent city areas that an agent is assumed to have heard of. For example, a resident may know about the Times Square area in NYC through public information without having visited it. We obtain these regions from public webpages, maps, and review websites. The resulting lists contain 33 districts in Beijing and 35 in NYC. All agents in the same city share this candidate list. However, we sample a subset separately for each agent, so individuals can begin the simulation knowing about different areas of the city. Each selected region is stored with a center and an approximate radius, describing where the area is and roughly how far it extends. It does not contain a list of places personally visited by the agent, since it represents general awareness of an area rather than knowledge built from visits. Visited regions are initially empty. They are formed during simulation by clustering the agent's visited places.

A.5 Cognitive Map Updating and Forgetting

Region-level updating. Visited regions represent areas where an agent's visited places are concentrated. At the end of each simulated day, we re-cluster all visited places currently retained in the cognitive map, excluding home and work anchors.

We use DBSCAN with haversine distance and a 500 m neighborhood radius. A place is a core point if its neighborhood contains at least three places, including itself. The resulting clusters define the updated visited regions and replace the previous grouping. Unclustered places remain in place-level memory. The 500 m threshold determines which places are neighbors, rather than fixing the size of a region. Connected groups of nearby places can therefore form a region extending beyond 500 m.

Each visited region stores its member places, their average coordinates as its center, and the distribution of place categories. It is named after a local district or nearby landmark. Its radius is the distance from the center to the farthest member place.

Memory decay. At the end of each simulated day, the memory strength of non-anchor places decays exponentially: $s_p \leftarrow s_p \exp(-\lambda_p \Delta t)$, where s_p is the memory strength of place p, λ_p is its decay rate, and Δt is the elapsed time in days. Daily updates use $\Delta t = 1$. Home and work are treated as stable anchors and are exempt from decay.

The decay rate depends on how often a place has been visited or how the agent learned about it. For visited places, we use $\lambda_p = 0.020$. Repeated visits reinforce memory strength, allowing frequently revisited places to remain strongly represented in memory for longer. For known-of places, we use $\lambda_p = 0.040$ for information obtained through search or perception, and $\lambda_p = 0.020$ for shared city

knowledge. The lower rate reflects the assumption that shared city knowledge is more stable and less easily forgotten. All rates are measured per day. These rates are fixed design choices.

A.6 Intent-guided Retrieval

Intent-guided retrieval constructs a bounded set of candidate places for the planned activity. It uses the agent's initial idea of where to go, with any preferences(i.e., distance and novelty) provided during planning, to determine where to search and which places to include.

Specific-place intents. An agent may think of an activity together with a particular place. For example, it may want to eat ramen at Ichiran, rather than simply decide to have lunch. In this case, retrieval looks for that restaurant in the agent's visited or known-of places using its name or identifier. If multiple branches share the same name, the nearest one is retrieved. If no matching place is found, retrieval searches more broadly for places suitable for the activity.

Region-level intents. When an agent has a region in mind, such as wanting coffee in Sanlitun, retrieval uses the corresponding region in its cognitive map to define where to search. Within this region, it looks for visited and known-of places suitable for the planned activity. Using the ranking procedure described below, it retains up to ten visited places and five known-of places. In addition to recalling places from memory, the agent can search for other options within the chosen region, modeling how people consult maps or apps when they know little about an area or want to try somewhere new.

Open intents. An agent may know what it wants to do without having a particular place or area in mind, such as wanting coffee but not knowing where to go. In this case, retrieval searches for suitable places around the agent's current location.

The search radius depends on the distance preference provided during planning. A preference for close, medium, or far corresponds to an initial radius of 1.5, 5, or 10 km, respectively. If no distance preference is specified, retrieval uses 20 km. Within this search scope, retrieval considers both remembered places and places outside the cognitive map. By default, the candidate set includes: (1) visited places, providing familiar options; (2) known-of places, providing opportunities to visit places the agent has heard of. (3) places from the external POI database that are not yet in the cognitive map, representing options discovered through search. If the agent wants to try somewhere new, retrieval reduces the visited-place and increases the known-of and external-place. Thus, more candidates are places the agent has not visited, while familiar alternatives remain available.

Ranking places in the cognitive map. Retrieval first identifies places within the search area that are suitable for the planned activity. It then ranks them to determine which places enter the candidate set. Visited places are ranked using three factors: memory strength, personal place appeal (how much the agent likes each place), and how closely the agent's usual visiting times for that place match the current time of day. For known-of places, priority depends on memory strength: places the agent remembers more clearly are more likely to enter the candidate set. The agent then chooses a destination from this set in the next stage.

The resulting candidates are annotated with their memory status, visit counts, distance, and estimated travel information before being passed to destination selection.

B Details of the Experimental Setup

B.1 Dataset

We use two public mobility datasets from Beijing and New York City (NYC).

Beijing. The Beijing dataset contains mobility trajectories and user profiles from approximately 100K users over a three-month period, from October 1, 2019 to December 31, 2019 Shao et al. (2026). Each trajectory consists of temporally ordered location records, and the associated user profiles include demographic attributes such as age, gender, and occupation. For our experiments, we simulate 1K users using their mobility context in November 2019. We use trajectories from an additional 20K real users randomly sampled from the same month as ground truth. The two user cohorts do not overlap.

New York City. The NYC dataset is a Foursquare POI check-in dataset containing records from approximately 1K users over about 10 months, from April, 2012 to February, 2013 Yang et al. (2015). Since the dataset does not provide demographic profiles for agent initialization, we construct synthetic profiles for 500 simulated users using U.S. Census population statistics U.S. Census Bureau (2014). Each synthetic profile includes sex, age, education, employment status, industry, and personal and household income where applicable. Attributes are jointly sampled from Census microdata calibrated to demographic and income distributions, with home locations assigned to census block groups. We simulate their mobility over June 2012 and use all real check-in records observed during the same month for evaluation.

Following prior mobility-generation studies JIAWEI et al. (2024); Shao et al. (2026), we evaluate simulated mobility against observed check-in records. These records exhibit several established human mobility regularities, as illustrated in Appendix E. However, check-ins capture only a subset of individuals' activities. The NYC evaluation therefore assesses whether the simulation reproduces the mobility patterns observed in these records, rather than complete daily mobility.

B.2 BASELINES

We compare our method with two categories of human mobility simulation baselines: *training-free LLM-based methods*, which directly leverage LLMs without task-specific model training, and *training-based methods*, which require additional training or fine-tuning on mobility data. All baselines are adapted to the same information-availability setting, while preserving their original generation mechanisms whenever applicable.

Training-free LLM-based baselines:

- **LLMob:** It JIAWEI et al. (2024) follows a two-phase framework. In Phase 1, it identifies a user persona by evaluating how well candidate persona descriptions explain the user's historical mobility trajectories. In Phase 2, mobility is generated day by day. For each day, the motivation inference and candidate-location retrieval are conditioned on the real mobility record from the previous day.

  Our adaptation: Since historical trajectories are unavailable in our setting, we use the raw demographic profile directly as the persona. During generation, we adopt the evolving-based setting, using the agent's most recently generated day as the reference. Candidate locations are retrieved from AMap POIs near the agent's home and workplace, with at most 20 candidates per category.

- **CoPB:** The original CoPB Shao et al. (2026) framework uses a lightweight fine-tuned LLaMA 3-8B model to generate daily activity schedules, including activity categories and their corresponding time intervals. It provides two approaches for grounding these activities into geographic locations: a gravity-based model and a diffusion-based model.

  Our adaptation: We adopt CoPB's gravity-based location model and retain the original implementation and parameter settings. The only modification is replacing the original POI database with our Gaode (Amap) POI database.

- **UML:** UML Bhandari et al. (2024) generates synthetic daily travel diaries using an LLM. It first samples user demographics from population statistics and then prompts the LLM to generate a sequence of daily activities, including visited location types and their arrival and departure times. The original method focuses on activity-level mobility generation and does not produce precise geographic locations.

  In our implementation, we keep the original activity-generation procedure unchanged. To obtain concrete locations required by our evaluation setting, we additionally apply a gravity model to map each generated activity to a candidate POI of the corresponding category.

- **CitySim:** CitySim Bougie & Watanabe (2025) uses LLM-driven agents with personas, needs, goals, and memories to recursively generate daily activities, followed by belief-aware gravity-based POI selection. Since no official source code is available, we reimplement its core mobility-generation ideas based on the methodology described in the paper and adapt them to our experimental setting.

Training-based baselines:

- **Geo-LLaMA.** Geo-LLaMA Li et al. (2024) fine-tunes an LLM on visit-permuted trajectories and enables controlled generation through visit-based prompts. We adapt this approach by LoRA-fine-tuning Llama-2-7B separately for each city, using 15-minute time slots. Generation is conditioned on available home/work or historical frequent-location anchors, with grammar-constrained decoding for structured outputs.
- **Movesim:** MoveSim Feng et al. (2020) is a GAN-based mobility simulator that generates sequences using self-attention, spatial relations, and mobility-aware adversarial training.

  In our implementation, we use MoveSim's core architecture and adversarial training framework, representing each day as 48 half-hour slots. For NYC, unobserved slots are filled with the most recent check-in location; slots before the first check-in are filled with that first location. The first simulated day starts from the available grid cell nearest to each user's home, and subsequent days start from the previous day's final cell. Consecutive slots in the same cell are merged into stay events with start and end timestamps.
- **MIRAGE:** MIRAGE Deng et al. (2025) combines an intensity-free neural temporal point process with neural exploration-and-preferential-return dynamics. A user VAE generates preference embeddings for synthetic individuals.

  In our implementation, we convert Beijing stay-point records into MIRAGE's input format, retaining the available coarse activity labels: Home, Work, and Other. For NYC, we retain the POI categories provided in the check-in data. For both datasets, we generate trajectories over the month-long evaluation period using VAE-sampled user embeddings.
- **TimeGeo:** TimeGeo Jiang et al. (2016) uses time-dependent probabilities to decide whether a person stays, moves, or returns home. It selects other destinations through rank-based exploration and preferential return (r-EPR), favoring nearby locations when exploring new places and frequently visited locations when returning to familiar places.

  In our implementation, we estimate mobility parameters from the training data. For Beijing, we use home and work locations from user profiles and transfer behavioral parameters from training users. For NYC, we infer home locations from historical check-ins. In both cities, each user's visitation history is preserved across simulation days.

B.3 VARIANTS

To disentangle the contribution of each component in our method, we conduct ablation experiments, each removing one mechanism from the full model, all else held fixed.

(1) CogMob w/o Cognitive Map-A (w/o CM-A). We remove the cognitive map and let the LLM generate each destination on its own, from no real-place pool at all.

(2) CogMob w/o Cognitive Map-B (w/o CM-B). We remove the cognitive map, and instead let the agent pick each destination from real POIs (by distance and rating). The pool is anchored to the agent's current location but has no memory of where this person has been before, so it cannot favor their familiar places.

(3) CogMob w/o Dynamic Retrieval (w/o DR). We remove the intent-conditioned retrieval and replace it with a fixed one: instead of choosing how to fetch candidates based on the agent's planned intent (recall a specific remembered place, focus on a target region, or search openly), every step just retrieves from the same nearby pool. The cognitive map and memory modules are kept.

B.4 IMPLEMENTATION

CogMob is implemented in Python 3.12 and uses Qwen3-32B-AWQ Yang et al. (2025) as its backbone LLM, with a single model instance deployed across two NVIDIA RTX A5000 GPUs. All training-free LLM-based baselines use the same backbone to control for differences in model capability. For CogMob, we simulate up to five agents concurrently, each generating its activities sequentially. Depending on the stage, temperatures range from 0.7 to 1.0, and maximum output lengths range from 300 to 1,200 tokens. We report the mean and standard deviation across three simulation runs. Region clustering is performed using DBSCAN from scikit-learn. POIs are obtained from the Amap API Amap for Beijing and OpenStreetMap OpenStreetMap for New York City. Agent initialization is described in Appendix A.1, and cognitive-map initialization is detailed in Appendix A.4.

B.5 METRICS

We evaluate the simulation from three perspectives. The evaluation metrics have been widely used in prior mobility simulation studies Cornacchia & Pappalardo (2021); Yuan et al. (2026); JIAWEI et al. (2024); Shao et al. (2026); Feng et al. (2020); Li et al. (2024). Many of these mobility patterns have been widely studied and repeatedly observed in real-world data Brockmann et al. (2006); Gonzalez et al. (2008); Song et al. (2010a).

(1) First, we examine whether simulated individuals reproduce the fundamental mobility characteristics observed in real people, including stay duration, jump distance, radius of gyration, and visitation entropy.

- **Stay Duration (Δt).** Stay duration Brockmann et al. (2006) measures how long an individual remains at a visited location, reflecting the temporal characteristics of human activities. Stay duration is also referred to as *dwell time* in other works. For each stay i, it is defined as $\Delta t_i = t_i^{\text{end}} - t_i^{\text{start}}$. We aggregate the durations of all valid stays and compute the JSD between the real and simulated stay-duration distributions.
- **Jump Distance (Δr).** Jump distance Gonzalez et al. (2008), also referred to as *travel distance*, measures the spatial displacement between two consecutive stays and characterizes the typical travel range of an individual. It is defined as

$$\Delta r_i = d_{\text{hav}}\left(\mathbf{r}_i, \mathbf{r}_{i+1}\right), \tag{1}$$

  where $\mathbf{r}_i$ and $\mathbf{r}_{i+1}$ denote two consecutive stay locations and d_{hav} is the haversine distance. We aggregate all consecutive jumps and compute the JSD between the real and simulated jump-distance distributions.
- **Radius of Gyration (r_g).** Radius of gyration Gonzalez et al. (2008) measures the characteristic spatial range of an individual's movements around their center of mass. It is calculated as the root-mean-square distance of visited locations from the duration-weighted center of mass:

$$r_g = \sqrt{\frac{\sum_i \Delta t_i d_{\text{hav}}^2(\mathbf{r}_i, \mathbf{r}_c)}{\sum_i \Delta t_i}}, \tag{2}$$

  where $\mathbf{r}_i$ and Δt_i denote the location and duration of stay i, respectively, and $\mathbf{r}_c$ is the duration-weighted center of mass. We calculate one r_g value for each individual and compute the JSD between the real and simulated distributions of r_g.
- **Visitation Entropy (H).** Visitation entropy measures the uncertainty of an individual's location choices based on visitation frequencies without considering their temporal order. This metric corresponds to the temporal-uncorrelated entropy defined in prior human mobility research Song et al. (2010b). It is defined as

$$H = -\sum_{l \in \mathcal{L}} p_l \log_2 p_l, \tag{3}$$

  where $\mathcal{L}$ is the set of locations visited by the individual and p_l is the fraction of visits to location l. A higher entropy indicates a more diverse and evenly distributed visitation pattern, whereas a lower entropy indicates that visits are concentrated at a few locations. We calculate one entropy value for each individual and compute the JSD between the real and simulated distributions.

(2) Second, we assess whether simulated individuals reproduce the evolving patterns of human mobility over time, including exploration, return, visitation frequency, and capacity growth.

- **Exploration.** Exploration Song et al. (2010a) describes the progressive discovery of new locations. Let $S(t)$ denote the cumulative number of distinct locations visited up to time t. Its temporal growth follows

$$S(t) \sim t^{\mu}, \tag{4}$$

  where μ is the exploration exponent. A larger μ indicates a faster rate of discovering new locations. We evaluate exploration using two complementary measures: the *individual-level JSD*, which compares the distributions of individually fitted exploration exponents μ

and thus captures heterogeneity in exploration tendencies, and the *curve-level JSD*, which compares the population-average exploration curves $S(t)$ and reflects the growth in the diversity of visited locations over time.

- **Preferential Return.** Return behavior Song et al. (2010a) characterizes the tendency to revisit previously visited locations. Let f_i denote the historical visitation frequency of location i. Under preferential return, the probability of selecting this location for the next return is proportional to its visitation frequency: $\Pi_i = f_i$. Following this principle, we further characterize the strength of preferential return using the generalized relationship

$$\Pi(f) \sim f^{\gamma}, \tag{5}$$

where γ is the return exponent and $\gamma = 1$ corresponds to linear preferential return. We evaluate return behavior using two complementary measures: the *individual-level JSD*, which compares the distributions of individually fitted return exponents γ, and the *curve-level JSD*, which compares the population-level return-probability curves $\Pi(f)$ over the same visitation-frequency bins.

- **Visitation Frequency.** Visitation frequency Song et al. (2010a) describes how visits are distributed across locations. Locations are ranked in descending order of visitation frequency, and the visitation frequency f_k of the k-th most visited location follows

$$f_k \sim k^{-\alpha}, \tag{6}$$

where α is the visitation-frequency exponent. A larger α indicates that visits are more concentrated at a small number of highly ranked locations. We evaluate visitation frequency using the *individual-level JSD* between the distributions of fitted exponents α and the *curve-level JSD* between the population-average rank-frequency curves over the same ranks.

- **Capacity Growth.** Capacity growth Alessandretti et al. (2018) characterizes how the set of repeatedly visited locations expands over time. A location is considered part of the mobility capacity after it has been visited at least twice. The capacity set at time t is defined as

$$C(t) = |\{l : n_l(t) \geq 2\}|, \tag{7}$$

where $n_l(t)$ denotes the cumulative number of visits to location l up to time t. We evaluate mobility-capacity growth using the *curve-level JSD*, which compares the population-average capacity-growth curves $\langle C(t) \rangle$ of the real and simulated trajectories over the same observation periods. A lower JSD indicates that the simulation more accurately reproduces the temporal growth of recurrently visited locations.

(3) Third, we evaluate urban spatial patterns by assessing whether the simulated data reproduces the spatial distributions of visits (LocFreq) and origin–destination flows (ODSim) Shao et al. (2026) observed in real-world trajectories.

- **Location-Frequency (LocFreq).** LocFreq measures the similarity between the aggregate spatial distributions of real and simulated activities. We partition the city into a regular grid of approximately 2 km cells $\{c_1, \ldots, c_G\}$. All stays are assigned to their corresponding grid cells. The visitation probability of cell c_g is defined as

$$p_g = \frac{n_g}{\sum_{j=1}^{G} n_j}, \tag{8}$$

where n_g is the number of stays in cell c_g. Let P^{real} and P^{sim} denote the resulting real and simulated visitation distributions. LocFreq is the JSD between the real and simulated visitation distributions.

- **OD Similarity (ODSim).** ODSim compares origin–destination flow distributions between real and simulated trajectories. We map each individual's stays to grid cells and merge consecutive stays within the same cell. Let $N(o, d)$ denote the total number of flows from cell o to cell d across all individuals. The normalized OD distribution is

$$P(o, d) = \frac{N(o, d)}{\sum_{o', d'} N(o', d')}. \tag{9}$$

Normalization enables comparison across datasets of different sizes. We report ODSim as the JSD between the real and simulated OD distributions, with lower values indicating greater similarity.

Table 4: Extended Overall performance on the Beijing and NYC datasets. Each cell reports the mean, with the standard deviation shown on the second line. For Explore, Return, and Visit-Freq, user-level and curve-level JSD measure differences in the distributions of user-level behavioral statistics and the shapes of aggregate behavioral curves, respectively. Lower values are better. The best mean values are shown in bold.

Dataset	Method	Mobility Characteristics				Mobility Dynamics							Urban Spatial Patterns	
		Radius	Jump	Duration	Entropy	Explore		Return		Visit-Freq		Capacity	LocFreq	OdSim
						User	Curve	User	Curve	User	Curve			
BJ	UML	0.0403 ±0.0104	0.0716 ±0.0043	0.2239 ±0.0005	0.4181 ±0.0321	0.2705 ±0.0321	0.0021 ±0.0006	0.2929 ±0.0578	0.0066 ±0.0036	0.5343 ±0.0785	0.0736 ±0.0088	0.0031 ±0.0004	0.1996 ±0.0038	0.3636 ±0.0035
	LLMob	0.0585 ±0.0004	0.1486 ±0.0145	0.2582 ±0.0310	0.3427 ±0.0314	0.1840 ±0.0257	0.0077 ±0.0033	0.1231 ±0.0033	0.0065 ±0.0008	0.1951 ±0.0481	0.0134 ±0.0027	0.0046 ±0.0011	0.2414 ±0.0248	0.4675 ±0.0052
	CoPB	0.1631 ±0.0009	0.0574 ±0.0012	0.2036 ±0.0022	0.2834 ±0.0045	0.1200 ±0.0048	0.0020 ±0.0001	0.0528 ±0.0093	0.0061 ±0.0017	0.1259 ±0.0046	0.0400 ±0.0006	0.0018 ±0.0001	0.3057 ±0.0018	0.4404 ±0.0022
	CitySim	0.0518 ±0.0114	0.0709 ±0.0115	0.3966 ±0.0006	0.3467 ±0.0711	0.1504 ±0.0697	0.0102 ±0.0048	0.0568 ±0.0470	0.0038 ±0.0012	0.1424 ±0.0844	0.0231 ±0.0114	0.0013 ±0.0003	0.2802 ±0.0024	0.4562 ±0.0154
	CogMob	**0.0214** ±0.0024	**0.0292** ±0.0003	**0.0378** ±0.0012	**0.1331** ±0.0127	**0.0271** ±0.0016	**0.0002** ±0.0001	**0.0295** ±0.0061	**0.0014** ±0.0004	**0.0911** ±0.0155	**0.0017** ±0.0001	**0.0003** ±0.0001	**0.1544** ±0.0051	**0.2943** ±0.0183
NYC	UML	0.3988 ±0.0055	0.1251 ±0.0058	0.3712 ±0.0026	0.5849 ±0.0099	0.2274 ±0.0043	0.0197 ±0.0001	0.1104 ±0.0131	0.0031 ±0.0004	0.2473 ±0.0136	0.0105 ±0.0016	0.0031 ±0.0004	0.1220 ±0.0068	0.3950 ±0.0080
	LLMob	0.3222 ±0.0146	0.2567 ±0.0028	0.3478 ±0.0473	0.5697 ±0.0796	0.3770 ±0.0043	**0.0133** ±0.0007	0.2420 ±0.0192	0.0041 ±0.0006	0.4773 ±0.0313	0.0109 ±0.0035	0.0165 ±0.0035	0.1107 ±0.0142	0.4470 ±0.0013
	CoPB	0.2727 ±0.0052	0.0894 ±0.0042	0.3239 ±0.0100	0.1680 ±0.0057	0.0320 ±0.0024	0.0220 ±0.0012	0.2201 ±0.0098	0.0061 ±0.0026	0.3209 ±0.0043	0.0307 ±0.0005	0.0026 ±0.0005	0.1212 ±0.0069	0.3939 ±0.0095
	CitySim	0.4186 ±0.0052	0.1234 ±0.0002	0.3462 ±0.0568	0.1445 ±0.0098	0.1248 ±0.0099	0.0224 ±0.0023	0.2227 ±0.0384	0.0059 ±0.0040	0.2346 ±0.0213	0.0305 ±0.0010	0.0083 ±0.0005	0.1265 ±0.0039	0.3924 ±0.0073
	CogMob	**0.0690** ±0.0148	**0.0469** ±0.0089	**0.2381** ±0.0122	**0.1319** ±0.0047	**0.0295** ±0.0034	0.0185 ±0.0007	**0.0396** ±0.0113	**0.0025** ±0.0006	**0.0919** ±0.0001	**0.0043** ±0.0003	**0.0023** ±0.0001	**0.1070** ±0.0048	**0.3582** ±0.0232

C Supplementary Results for Main Experiments

This section provides supplementary results and analyses for the experiments in the main text.

C.1 Extended Overall Performance Evaluation

To complement the main evaluation, Table 4 presents an extended performance comparison on the Beijing and NYC datasets. This analysis includes additional metrics: user-level and curve-level JSD for exploration, return, and visitation-frequency behavior. These metrics assess the distributions of user-level behavioral statistics and the shapes of aggregate behavioral curves, respectively, providing complementary views of mobility patterns. We also report the mean and standard deviation of each metric to characterize performance variability across runs.

C.2 Extended Comparison with Training-Based Methods

This section provides detailed experimental settings and extended results for the comparison with training-based methods. We further investigate whether occasional long-distance travel contributes to the baselines' errors in jump distance and radius of gyration in Beijing.

Experimental setup. For Beijing, the baselines are trained on over 20,000 November trajectories from 1,000 individuals. The trained models then generate November trajectories for a separate cohort of 1,000 individuals. For NYC, the baselines are trained on real-world trajectories recorded before June and generate June trajectories for the same individuals. Because the dataset contains relatively few users, we separate training and evaluation by time but do not evaluate generalization to unseen individuals.

Extended comparison. Table 5 presents the extended comparison with additional metrics: user-level and curve-level JSD for exploration, return, and visitation-frequency behavior. These metrics assess the distributions of user-level behavioral statistics and the shapes of aggregate behavioral curves, respectively, providing complementary views of mobility patterns. We report the mean and standard deviation of each metric across runs to characterize performance variability.

Table 5: Extended comparison with training-based methods on Beijing and NYC. Each cell reports the mean, with the standard deviation on the second line. For Explore, Return, and Visit-Freq, user-level and curve-level JSD measure differences in the distributions of user-level behavioral statistics and the shapes of aggregate behavioral curves, respectively. Lower values are better, and the best mean values are shown in bold.

Dataset	Method	Mobility Characteristics				Mobility Dynamics							Urban Spatial Patterns	
		Radius	Jump	Duration	Entropy	Explore		Return		Visit-Freq		Capacity	LocFreq	OdSim
						User	Curve	User	Curve	User	Curve			
BJ	Geo-LLaMA	0.3115 ±0.0153	0.1538 ±0.0020	0.1264 ±0.0038	0.1587 ±0.0117	0.0668 ±0.0141	0.0012 ±0.0001	0.0813 ±0.0142	0.0263 ±0.0261	0.1023 ±0.0094	0.0150 ±0.0001	0.0022 ±0.0002	0.2210 ±0.0153	0.4359 ±0.0229
	Movesim	0.1482 ±0.0156	0.2317 ±0.0651	0.1595 ±0.0190	**0.0783** ±0.0329	0.1613 ±0.0344	0.0137 ±0.0027	0.1043 ±0.0357	0.0173 ±0.0123	0.2738 ±0.1241	0.0354 ±0.0144	0.0147 ±0.0098	0.3924 ±0.0131	0.7310 ±0.0304
	MIRAGE	0.3390 ±0.0166	0.4222 ±0.0013	0.1032 ±0.0071	0.3419 ±0.0278	0.0809 ±0.0108	0.0005 ±0.0001	0.0804 ±0.0241	0.0232 ±0.0071	0.1878 ±0.0173	0.0145 ±0.0030	0.0019 ±0.0003	0.1333 ±0.0013	0.6486 ±0.0074
	TimeGeo	0.0808 ±0.0055	0.1897 ±0.0009	0.0782 ±0.0022	0.2995 ±0.0055	0.0919 ±0.0085	0.0015 ±0.0000	0.0378 ±0.0060	0.0188 ±0.0176	0.1715 ±0.0098	0.0132 ±0.0007	0.0027 ±0.0001	**0.1049** ±0.0022	0.4476 ±0.0030
	CogMob	**0.0214** ±0.0024	**0.0292** ±0.0003	**0.0378** ±0.0012	0.1331 ±0.0127	**0.0271** ±0.0016	**0.0002** ±0.0001	**0.0295** ±0.0061	**0.0014** ±0.0004	**0.0911** ±0.0155	**0.0017** ±0.0001	**0.0003** ±0.0001	0.1544 ±0.0051	**0.2943** ±0.0183
NYC	Geo-LLaMA	0.1599 ±0.0074	0.0856 ±0.0115	0.4279 ±0.0023	0.3654 ±0.0149	0.1659 ±0.0431	0.0228 ±0.0004	0.2193 ±0.0188	0.0146 ±0.0063	0.1874 ±0.0289	0.0179 ±0.0021	0.0032 ±0.0002	0.1917 ±0.0042	0.6295 ±0.0031
	Movesim	0.2524 ±0.0993	0.0711 ±0.0067	0.4107 ±0.0487	0.3869 ±0.1803	0.2238 ±0.0634	0.0207 ±0.0076	0.2015 ±0.0259	0.0076 ±0.0058	0.3260 ±0.0723	0.0082 ±0.0022	0.0101 ±0.0115	0.2115 ±0.0241	0.6016 ±0.0337
	MIRAGE	0.1785 ±0.0194	0.0986 ±0.0053	0.2439 ±0.0025	0.1425 ±0.0141	0.1040 ±0.0218	0.0186 ±0.0015	0.0796 ±0.0082	0.0045 ±0.0014	0.1012 ±0.0072	0.0106 ±0.0031	0.0040 ±0.0008	0.0325 ±0.0036	0.3651 ±0.0078
	TimeGeo	**0.0366** ±0.0033	**0.0116** ±0.0012	0.3313 ±0.0019	0.2851 ±0.0168	0.0697 ±0.0146	0.0192 ±0.0049	0.2041 ±0.0174	0.0135 ±0.0023	0.1299 ±0.0026	0.0233 ±0.0009	0.0281 ±0.0023	**0.0239** ±0.0001	**0.3208** ±0.0058
	CogMob	0.0690 ±0.0148	0.0469 ±0.0089	**0.2381** ±0.0122	**0.1319** ±0.0047	**0.0295** ±0.0034	**0.0185** ±0.0007	**0.0396** ±0.0113	**0.0025** ±0.0006	**0.0919** ±0.0001	**0.0043** ±0.0003	**0.0023** ±0.0001	0.1070 ±0.0048	0.3582 ±0.0232

C.2.1 Reproducing Local and Long-Distance Mobility in Beijing

MIRAGE and TimeGeo overestimate both the median jump distance and the median radius of gyration in Beijing. The real-world trajectories of Beijing contain routine local trips alongside occasional long-distance intercity travel. To investigate whether these infrequent long-distance trips contribute to the observed errors, we compare results with and without out-of-city activity records.

We consider two settings: one retains all activity records in the Beijing dataset (i.e., unfiltered), including those located outside the city, while the other excludes records outside Beijing and focuses on within-city mobility (i.e., filtered). In each setting, we compare the corresponding real-data distributions with those generated by CogMob, MIRAGE, and TimeGeo. In the filtered setting, we first remove activity records outside Beijing from the real data and use the resulting dataset to train MIRAGE and fit TimeGeo. We apply the same spatial filter to the trajectories generated by CogMob. The filtered real data then serves as the ground truth for evaluating all three methods.

Figure 8 presents three complementary comparisons: (1) the distribution of jump distances, (2) the fraction of individuals whose maximum jump exceeds a given distance, and (3) the distribution of time-weighted radii of gyration. Solid and dashed lines denote the unfiltered and filtered settings, respectively. From the figure, we can see that without filtering, MIRAGE and TimeGeo overestimate median jump distances and radius of gyration, and generate long-distance movements for too many individuals. Their errors therefore affect both typical mobility and the tails. Filtering suppresses the long-distance tails, but the baselines' medians remain above those of the filtered ground truth. In both settings, CogMob more closely matches the corresponding observed distributions.

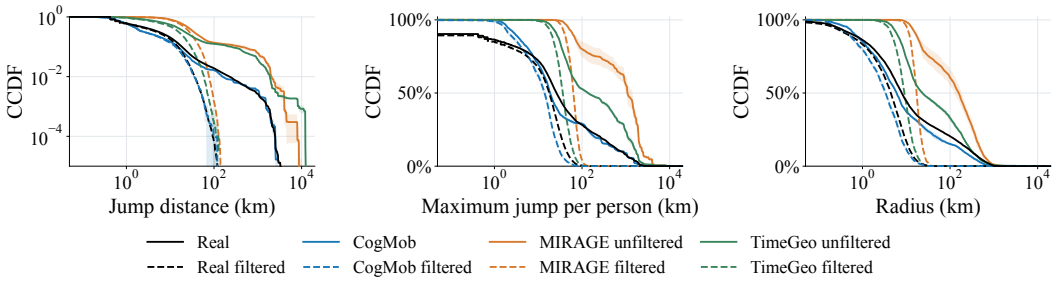


Figure 8: Effect of spatial filtering on simulated mobility in Beijing: CCDFs of jump distance (left), maximum jump per individual (middle), and time-weighted radius of gyration (right). Solid/dashed lines denote unfiltered/filtered settings; shaded bands indicate ±1 SD across runs.

These results show that MIRAGE and TimeGeo struggle to jointly reproduce routine local mobility and occasional long-distance travel. In contrast, CogMob more closely reproduces both typical mobility and the long-distance tails. Long-tailed travel-distance distributions have been documented in human mobility research Brockmann et al. (2006). Accurately capturing both routine local trips and rare long-distance travel is relevant to transportation planning Chen et al. (2023) and epidemic simulation and intervention analysis Lai et al. (2020).

C.3 Extended Ablation Study

Table 6 presents extended ablation results, including user-level and curve-level JSD for exploration, return, and visitation-frequency behavior. These metrics assess discrepancies in the distributions of user-level behavioral statistics and the shapes of aggregate behavioral curves, respectively, providing complementary views of how individual components affect simulation performance. We report the mean and standard deviation of each metric across runs to characterize performance variability.

Table 6: Extended ablation results on the Beijing dataset. Each cell reports the mean, with the standard deviation on the second line. For Explore, Return, and Visit-Freq, user-level and curve-level JSD measure differences in the distributions of user-level behavioral statistics and the shapes of aggregate behavioral curves, respectively. Lower values are better, and the best mean values are shown in bold.

Method	Mobility Characteristics				Mobility Dynamics							Urban Spatial Patterns	
					Explore		Return		Visit-Freq		Capacity		
	Radius	Jump	Duration	Entropy	User	Curve	User	Curve	User	Curve		LocFreq	OdSim
CogMob w/o CM-A	0.0322 ±0.0018	0.1278 ±0.0027	**0.0327** ±0.0002	0.1983 ±0.0112	0.2523 ±0.0221	0.0061 ±0.0010	0.6114 ±0.0292	0.0209 ±0.0103	0.7636 ±0.0187	0.0926 ±0.0001	0.0027 ±0.0001	0.4377 ±0.0064	0.6210 ±0.0095
CogMob w/o CM-B	0.0264 ±0.0066	0.0372 ±0.0001	0.0462 ±0.0007	0.1847 ±0.0020	0.1001 ±0.0005	0.0044 ±0.0007	0.1213 ±0.0070	0.0050 ±0.0052	0.1581 ±0.0058	0.0121 ±0.0007	0.0004 ±0.0001	0.1527 ±0.0074	**0.2935** ±0.0008
CogMob w/o DR	0.0291 ±0.0025	0.0384 ±0.0003	0.0400 ±0.0010	0.1677 ±0.0105	0.0326 ±0.0022	0.0004 ±0.0000	0.0302 ±0.0099	0.0021 ±0.0001	0.1050 ±0.0036	0.0023 ±0.0001	0.0004 ±0.0002	**0.1506** ±0.0081	0.3006 ±0.0137
CogMob	**0.0214** ±0.0024	**0.0292** ±0.0003	0.0378 ±0.0012	**0.1331** ±0.0127	**0.0271** ±0.0016	**0.0002** ±0.0001	**0.0295** ±0.0061	**0.0014** ±0.0004	**0.0911** ±0.0155	**0.0017** ±0.0001	**0.0003** ±0.0001	0.1544 ±0.0051	0.2943 ±0.0183

C.4 Extended Applications of Simulated Mobility Data

This section provides additional experimental details and results for the downstream applications of simulated mobility data. We supplement the NYC segregation analysis presented in the main text with results for Beijing.

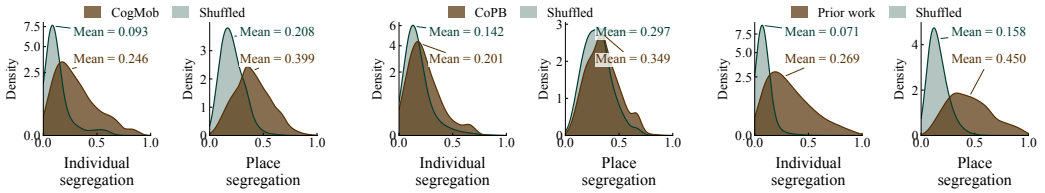
C.4.1 Socioeconomic Segregation

In this section, we examine whether similar segregation patterns emerge in simulated data by two questions: **Q1:** Do simulated mobility data exhibit non-random socioeconomic segregation at both the individual and place levels? **Q2:** Does the association between exploration behavior and individual experienced segregation reported in prior work also appear in simulated mobility data? We conduct these experiments using simulated mobility data from both Beijing and NYC.

Experimental setup. **NYC:** We analyze simulated trajectories from 4,000 individuals across NYC's five boroughs. We use home block-group median household income from the 2009–2013 ACS U.S. Census Bureau (2014) as a proxy for wealth and assign individuals to four groups using citywide household-weighted quartiles. Following the original study Moro et al. (2021), we retain stays lasting 5 minutes to 1 day. We randomly shuffle visit destinations 30 times, keeping each individual's visit counts and durations unchanged, and compare segregation with the simulation.

Beijing: We analyze simulated trajectories from 3,000 individuals. Socioeconomic groups are defined using consumption categories, with the two highest categories combined to form four groups. We retain stays lasting 5 minutes to 1 day. We use the same randomization and correlation procedures as for NYC. For comparison within LLM-based mobility simulation, we select CoPB Shao et al. (2026) for its strong overall performance. We use simulated trajectories from approximately 3,000 individuals per method.

For quantitative comparisons, we use JSD to compare distributions from simulated trajectories with those from the pooled public data of 11 cities Moro et al. (2021). (1) For Q1, we compare the distributions of individual experienced segregation and place segregation between simulated and



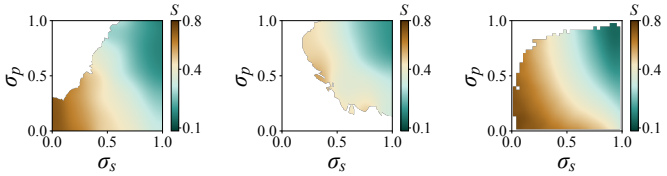
(a) CogMob: Beijing (b) CoPB: Beijing (c) Prior work

Figure 9: Individual experienced segregation and place segregation for CogMob and CoPB in Beijing, and data from 11 cities in prior work Moro et al. (2021). Each panel compares segregation distributions from the original visits with those obtained after randomizing visit destinations.

real data. (2) For Q2, we use three-dimensional JSD to compare the joint distribution of place exploration, social exploration, and individual experienced segregation between simulated users and real data. Using individual-level values, we divide each variable into 10 equal-width intervals and compute JSD between the proportions of users in the resulting $10 \times 10 \times 10$ bins. Lower JSD indicates that the simulated data more closely match how exploration behavior and segregation are jointly distributed in the real data.

Results of Beijing (1) Q1: Non-random socioeconomic segregation. We first examine whether the simulated Beijing trajectories exhibit socioeconomic segregation beyond that observed under randomized destination choices. Figure 9 compares segregation distributions for CogMob (a) and CoPB (b) in Beijing, and pooled data from 11 cities in prior work (c) Moro et al. (2021). Each panel shows individual experienced segregation (left) and place segregation (right), comparing original vs. shuffled visits. For CogMob, segregation varies across individuals and places in the simulated Beijing data. Both individual experienced segregation and place segregation are higher overall than after randomly shuffling visit destinations, indicating non-random socioeconomic segregation in our simulation. For a quantitative comparison, Table 7 reports JSD between the simulated segregation distributions and those from the pooled public data of 11 cities Moro et al. (2021). CogMob achieves JSD values of 0.0083 for individual experienced segregation and 0.0289 for place segregation, lower than those of CoPB, indicating closer segregation distributions to the reference data.

(2) Q2: Exploration and individual experienced segregation. Figure 10 shows mean individual experienced segregation for CogMob (a) and CoPB (b) in Beijing, and combined data from 11 cities in prior work (c) Moro et al. (2021). The horizontal and vertical axes represent social exploration and place exploration, respectively, while color indicates mean segregation. Given the limited number of simulated individuals, we use Gaussian smoothing to estimate local mean segregation and obtain a continuous visualization. The results show that higher levels of both place exploration and social exploration are generally associated with lower mean individual experienced segregation in the simulated Beijing data, consistent with prior findings Moro et al. (2021).



(a) CogMob (b) CoPB (c) Prior work

Figure 10: Exploration and individual experienced segregation for CogMob and CoPB in Beijing, and prior work Moro et al. (2021). Color indicates mean segregation.

Table 7: JSD between distributions from simulated Beijing trajectories and those from the pooled public data of 11 cities Moro et al. (2021). Lower values indicate more similar distributions.

Method	JSD ↓		
	Individual segregation	Place segregation	Exploration–segregation (3D)
CoPB	0.0307	0.0475	0.6245
CogMob	0.0083	0.0289	0.1111

For a quantitative comparison, Table 7 reports three-dimensional JSD for the joint distribution of the two exploration measures and individual experienced segregation. The quantitative comparisons use the original individual-level values without smoothing. CogMob achieves a JSD of 0.1111, lower than CoPB (0.6245), indicating that its joint distribution more closely matches the pooled public data from 11 cities.

D ADDITIONAL EXPERIMENTS

This section presents additional experiments that investigate aspects of CogMob beyond the scope of the main experiments.

D.1 DIFFERENT LLM BACKBONES

CogMob uses an LLM as the reasoning backbone of each mobility agent. To assess how backbone choice affects simulation performance, we evaluate Qwen3-8B Yang et al. (2025), Llama-3.1-8B-Instruct Grattafiori et al. (2024), and Qwen3-32B-AWQ by simulating the same 500 individuals on the Beijing dataset. The cognitive-map mechanism, prompts, retrieval procedure, and simulation settings are held fixed across all models.

As shown in Table 8, (1) Qwen3-32B-AWQ achieves the lowest mean JSD on eight of the thirteen metrics, with advantages spanning mobility characteristics, mobility dynamics, and urban spatial patterns. This supports its use as the default backbone in our main experiments, although no model performs best on every metric. (2) Qwen3-8B performs worse than Qwen3-32B-AWQ on twelve of the thirteen metrics, with particularly large gaps in mobility dynamics metrics. These results show that backbone choice substantially affects simulation quality even when the remaining framework components are held fixed. During simulation, we also observed cases where Qwen3-8B's reasoning was inconsistent with its final decisions. Such inconsistencies may contribute to its weaker performance. (3) However, parameter count alone does not determine simulation quality. Llama-3.1-8B-Instruct outperforms Qwen3-8B on ten of the thirteen metrics and achieves the best results among all three backbones on four metrics. These results demonstrate the potential of smaller models for mobility simulation and highlight the importance of model choice beyond parameter scale.

Table 8: Performance of CogMob with different LLM backbones on the Beijing dataset.

Method	Mobility Characteristics				Mobility Dynamics							Urban Spatial Patterns	
	Radius	Jump	Duration	Entropy	Explore		Return		Visit-Freq		Capacity	LocFreq	OdSim
					User	Curve	User	Curve	User	Curve			
Qwen3-8B	0.0653 ±0.0006	0.0478 ±0.0029	**0.0299** ±0.0049	0.2726 ±0.0534	0.1732 ±0.0258	0.0104 ±0.0016	0.2830 ±0.0085	0.0023 ±0.0005	0.4518 ±0.0196	0.0161 ±0.0044	0.0094 ±0.0011	0.2567 ±0.0026	0.4083 ±0.0055
Llama3.1-8B	0.1230 ±0.0072	0.0431 ±0.0018	0.0585 ±0.0042	**0.1222** ±0.0019	0.0368 ±0.0039	0.0020 ±0.0003	**0.0221** ±0.0002	0.0028 ±0.0000	**0.0603** ±0.0013	0.0031 ±0.0001	0.0013 ±0.0003	0.1808 ±0.0139	**0.3235** ±0.0132
Qwen3-32B-AWQ	**0.0307** ±0.0006	**0.0332** ±0.0019	0.0390 ±0.0007	0.1482 ±0.0016	**0.0315** ±0.0018	**0.0002** ±0.0000	0.0358 ±0.0006	**0.0022** ±0.0002	0.1023 ±0.0074	**0.0016** ±0.0002	**0.0003** ±0.0001	**0.1756** ±0.0002	0.3404 ±0.0072

D.2 SOCIAL INFORMATION INTERVENTIONS FOR MOBILITY REALISM

We investigate whether providing agents with additional external information improves mobility simulation. In real-world settings, individuals obtain information about places from map applications, popularity rankings, and social media, supplementing their existing knowledge of the city and informing their mobility decisions. We introduce such information through two social-information interventions and evaluate their effects on mobility realism, particularly the spatial distribution of visits and the reproduction of citywide hotspots.

We introduce two social-information interventions. (1) The popularity-signal intervention: In the simulation, agents observe place popularity during searches and retain it in their cognitive maps, modeling how people learn about and remember popular places through map and recommendation apps. (2) The social-media recommendation intervention exposes agents to random subsets of historically popular venues in several temporal waves. Recommended places are temporarily pinned in their cognitive maps and may influence subsequent planning and destination choices, modeling how people encounter recommendations and choose whether to follow them. We compare three configurations: CogMob, CogMob w Pop, and CogMob w Push. CogMob serves as the baseline, while CogMob w Pop adds popularity signals and CogMob w Push adds social-media recommendations. Each configuration simulates 500 individuals in Beijing during November. Place-popularity information is obtained from October mobility records outside the National Day holiday period.

The results reveal three main findings. **First, external social information improves urban spatial realism.** We evaluate urban spatial patterns using LocFreq and ODSim, and assess hotspot identification and ranking using P@40 and NDCG@40, respectively.

P@40 measures the overlap between the 40 most-visited grid cells in simulated and real data. NDCG@40 evaluates the simulated ranking using real visit counts, giving greater weight to higher-ranked cells. Higher values indicate better performance for both metrics. As shown in Figure 11, both interventions increase P@40 and NDCG@40 while reducing LocFreq and OdSim divergence, indicating improvements in hotspot identification and ranking, as well as the spatial distributions of visits and origin–destination flows.

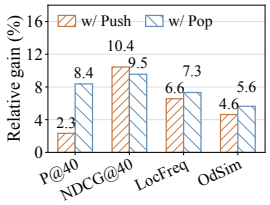


Figure 11: Relative gain over CogMob from popularity (w/ Pop) and social-media recommendations (w/ Push).

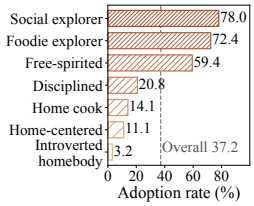


Figure 12: Adoption rates of social-media recommendations across persona archetypes.

Second, gains in urban spatial realism coexist with improvements in other mobility dimensions. Incorporating social information also improves several metrics of mobility characteristics and dynamics, suggesting that its benefits extend beyond citywide spatial patterns. However, these improvements involve some trade-offs across metrics. Detailed results are provided in Table 9.

Third, agents respond heterogeneously rather than uniformly to social recommendations. For the social-media recommendation intervention, we examine whether agents follow recommendations automatically or selectively act on them based on their own preferences and current context. We define the adoption rate as the fraction of exposed agents who visit at least one recommended POI after receiving the recommendation, and compare this rate across persona archetypes. As shown in Figure 12, adoption ranges from 3.2% among the most introverted agents to 78.0% among social explorers, a more than 24-fold difference. These results show that agents do not follow recommendations uniformly and that their responses vary with individual behavioral characteristics.

Table 9: Effect of external information interventions on the Beijing dataset. Explore, Return, and Visit-Freq are evaluated using both user-level and curve-level JSD. Lower values indicate better results.

Method	Mobility Characteristics				Mobility Dynamics						
	Radius	Jump	Duration	Entropy	Explore		Return		Visit-Freq		Capacity
					User	Curve	User	Curve	User	Curve	
CogMob w/ Pop	**0.0264**	0.0291	0.0397	**0.1458**	**0.0248**	**0.0002**	0.0443	**0.0013**	0.1112	**0.0015**	**0.0002**
CogMob w/ Push	0.0415	**0.0288**	0.0408	0.1563	0.0268	**0.0002**	0.0433	0.0019	0.1061	0.0022	0.0004
CogMob	0.0307	0.0332	**0.0390**	0.1482	0.0315	**0.0002**	**0.0358**	0.0022	**0.1023**	0.0016	0.0003

D.3 Qualitative analysis of individual mobility footprints

Different groups of people tend to have different spatial patterns of activity. For example, some visit places across a wide area, while others concentrate their activities around home and work. We compare real and simulated mobility footprints to examine whether CogMob captures these differences and how these patterns change when the cognitive map is removed.

We classify users into four groups by mobility style (Explorer or Homebody) and commuting distance (long or short). Mobility style is determined using a shared clustering model based on radius of gyration, exploration and return exponents, visitation-frequency exponent, and location entropy. The same model is applied to real and simulated users, while commuting distance is measured between home and work. We select two users from each group in the real and CogMob populations to compare their mobility footprints. For w/o CM-A, we use the same agents and personas as in CogMob, but remove the cognitive map. We keep each agent in the same group and compare its mobility footprints under the two settings.

Figure 13 shows similar differences between mobility groups in the real data and CogMob: Homebody users concentrate their activities around home and work, while Explorer users visit places across broader areas. Compared with CogMob, Homebody agents in w/o CM-A visit places across a wider area, even though their personas remain unchanged. Their footprints become more similar to those of Explorer agents. These examples suggest that personalized cognitive maps help reproduce individual differences in spatial mobility patterns.

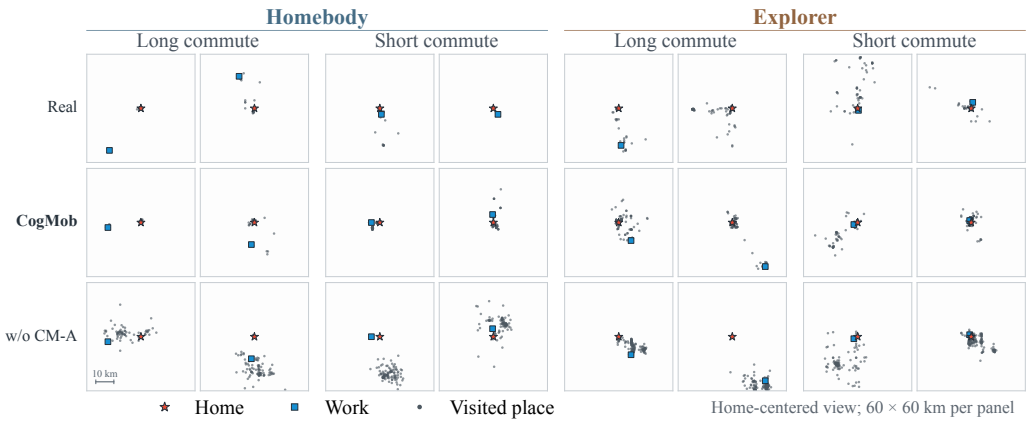


Figure 13: Qualitative comparison of individual mobility footprints grouped by mobility style (Homebody or Explorer) and commuting distance (long or short). Rows show real data, CogMob, and w/o CM-A, with two examples per group in each row.

D.4 A Simple Baseline with Visitation History

We examine whether accumulated visitation history provides sufficient spatial memory for mobility simulation. We construct a baseline, LLM w/ Visit History, that generates daily activities and destinations from an individual's profile and an accumulated history of previously simulated visits. The history is maintained as a flat list of visited locations with their cumulative visit counts and is updated throughout the simulation.

As shown in Table 10, CogMob achieves lower JSD across all evaluated metrics, with particularly large improvements in exploration, return behavior, and location-capacity growth. These results suggest that accumulated visitation history alone is insufficient to reproduce mobility patterns as accurately as CogMob, which explicitly represents personalized spatial knowledge and uses it to guide retrieval and destination selection.

Table 10: Evaluation of visitation history as spatial memory on the Beijing dataset.

Method	Mobility Characteristics				Mobility Dynamics							Urban Spatial Patterns	
	Radius	Jump	Duration	Entropy	Explore		Return		Visit-Freq		Capacity	LocFreq	OdSim
					User	Curve	User	Curve	User	Curve			
LLM w/ Visit History	0.0712	0.1457	0.1966	0.3024	0.1876	0.0129	0.3570	0.0586	0.2588	0.0204	0.0130	0.1933	0.4833
CogMob	**0.0214**	**0.0292**	**0.0378**	**0.1331**	**0.0271**	**0.0002**	**0.0295**	**0.0014**	**0.0911**	**0.0017**	**0.0003**	**0.1544**	**0.2943**

E Visualization Results

This section visually examines whether the simulated mobility data reproduce the scaling laws of human mobility and other patterns observed in real-world data. We evaluate the simulation from three complementary perspectives. First, we compare fundamental mobility characteristics, including stay duration, jump distance, and radius of gyration. Second, we examine mobility dynamics, including exploration growth, return behavior, visitation frequency, and mobility-capacity growth, to assess patterns of discovering and revisiting locations over time. Third, we assess population-level urban spatial patterns by comparing the spatial distribution of visits between simulated and real-world data.

E.1 Fundamental Mobility Characteristics

We compare the complementary cumulative distribution functions (CCDFs) of jump distance, radius of gyration, and stay duration between simulated and real-world data in Beijing and NYC (Figures 14 and 15). These metrics characterize movement between consecutive locations, the spatial extent of individual mobility, and the time spent at a location, respectively.

In both cities, CogMob closely reproduces the distributions of jump distance and radius of gyration, capturing their overall shapes and decay patterns. The stay-duration distributions show larger discrepancies, particularly in their tails, with simulated data containing a lower proportion of long

stays. This gap is especially pronounced in NYC and may partly reflect differences in how durations are measured. In the real POI check-in data, intervals between consecutive check-ins may include unobserved activities and do not necessarily represent continuous stays at a single location. By contrast, simulated durations are derived from explicitly generated activity sequences. This difference limits the direct comparability of the real and simulated stay-duration distributions.

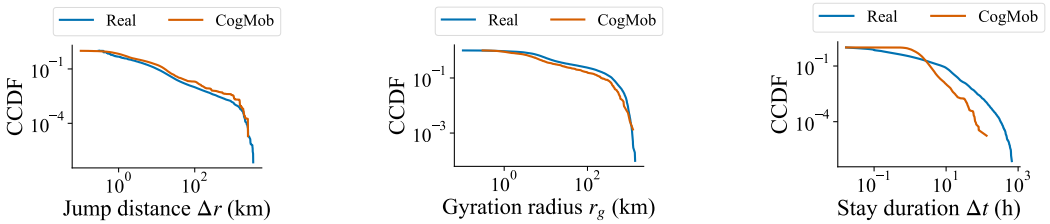


Figure 14: Beijing: CCDFs of jump distances, radii of gyration, and stay durations.

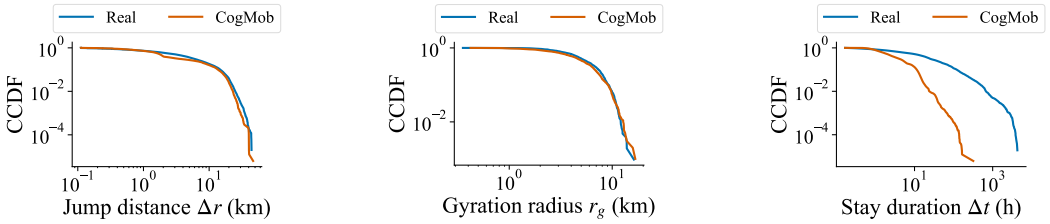
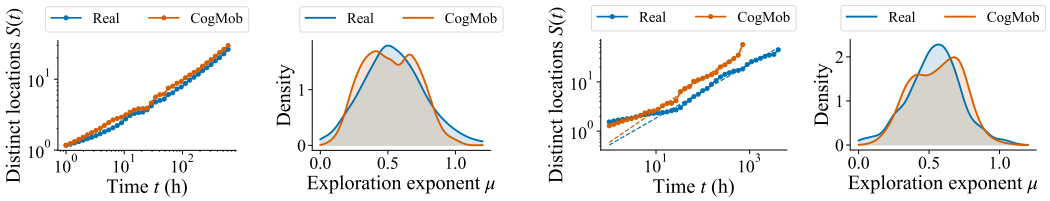


Figure 15: NYC: CCDFs of jump distances, radii of gyration, and stay durations.

E.2 Mobility Dynamics

We next examine mobility dynamics, focusing on how individuals explore new locations and return to familiar ones over time, how their location capacity grows, and how repeated visits accumulate into visitation-frequency patterns.

Exploration and referential return: Figures 16 and 17 examine two complementary aspects of mobility: exploring new locations and returning to familiar ones. For exploration, CogMob closely reproduces the growth in the number of distinct visited locations, $S(t)$, in Beijing, but shows faster growth in NYC over much of the shared observation period. The distributions of individual exploration exponents μ broadly overlap between real and simulated data in both cities, suggesting that CogMob captures variation in exploration behavior across individuals. For preferential return, CogMob captures the increase in return probability with past visitation frequency, reproducing the tendency to revisit frequently visited locations. The distributions of individual return exponents γ also closely align, suggesting that CogMob captures both the aggregate return pattern and individual differences in return behavior.

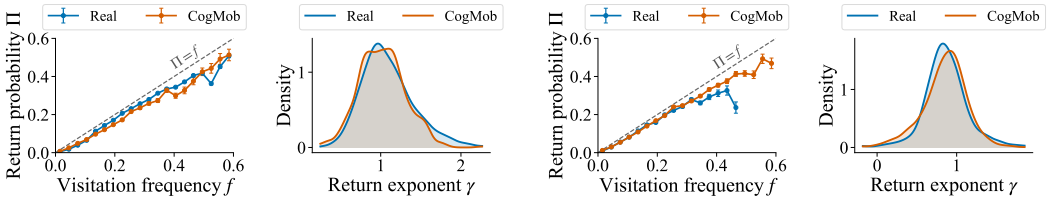


(a) Beijing (b) NYC

Figure 16: Exploration in real and CogMob-generated trajectories. For each city, the left panel shows the growth in the number of distinct visited locations $S(t)$ over time, and the right panel shows the distributions of individual exploration exponents μ.

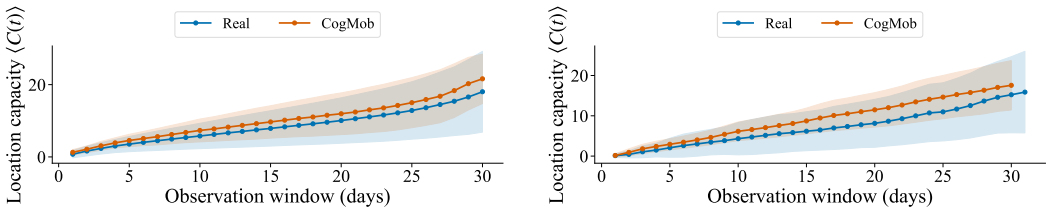
Location-capacity growth: Figure 18 examines location-capacity growth. CogMob captures the increase in location capacity with longer observation windows in both cities, although differences remain in capacity levels.

Visitation frequency: Figure 19 examines the distribution of visits across locations. The simulated rank-frequency curves closely follow those of the real data, capturing the concentration of visits at



(a) Beijing (b) NYC

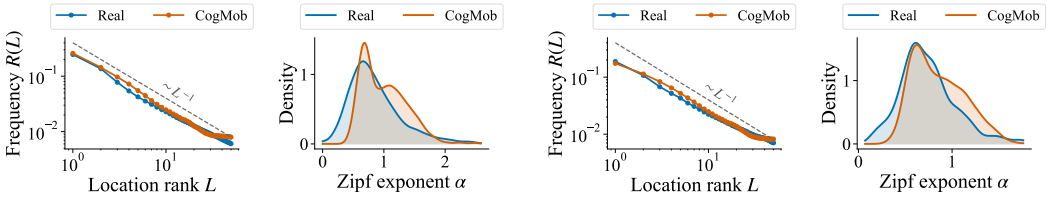
Figure 17: Preferential return behavior in real and CogMob-generated trajectories. For each city, the left panel shows return probability Π as a function of visitation frequency f, and the right panel shows the distributions of individual return exponents γ.



(a) Beijing (b) NYC

Figure 18: Growth of location capacity over time for real and CogMob-generated trajectories. Shading denotes one standard deviation.

a small number of frequently visited locations. The distributions of individual visitation-frequency exponents α also broadly overlap, suggesting that CogMob captures individual differences in how strongly visits are concentrated across locations. However, discrepancies remain in the peaks and shapes of these distributions.



(a) Beijing (b) NYC

Figure 19: Visitation-frequency patterns in real and CogMob-generated trajectories. For each city, the left panel shows rank-frequency curves, and the right panel shows distributions of exponents α.

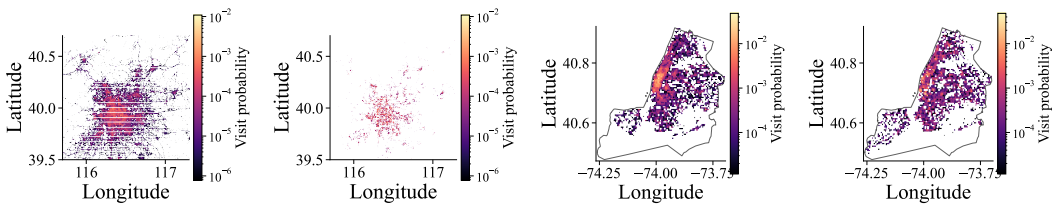
E.3 Urban Spatial Patterns

Finally, we examine whether simulated trajectories collectively reproduce the spatial patterns of urban city. Figure 20 compares visit probabilities aggregated into 500 m grid cells in Beijing and NYC. In Beijing, CogMob captures the broad concentration of visits in the central urban area, although the simulated map is spatially sparser, particularly in peripheral areas. This difference is partly attributable to the substantially smaller simulated population, which provides fewer observations of less frequently visited locations. In NYC, the simulated and real-data maps appear more similar, with comparable concentrations of activity along Manhattan and broadly similar spatial coverage.

F Prompt

F.1 Planning Prompt

Given the agent's persona, need state, episodic memory, and a region-level view of its cognitive map, this prompt decides the next activity and a coarse spatial intent—a specific place, a familiar region, or an open intent—which the retrieval stage then grounds into candidates.



(a) Beijing: Real (b) Beijing: CogMob (c) NYC: Real (d) NYC: CogMob

Figure 20: Spatial distributions of visit probabilities in real and CogMob-generated trajectories in Beijing and NYC.

Stage 1: planning prompt (Abridged example)

System: You are a resident of Beijing living out an ordinary day. Given who you are, how you feel right now, what you still owe or have planned, and the places you know, decide what to do next. Think like this specific person, not a generic optimizer. Output only JSON.

[Who you are]

Female, 29, marketing specialist at a media firm in the CBD. A dedicated foodie who seeks out specialty restaurants and quiet independent cafés; socially selective (a few close friends rather than large gatherings); enjoys exploring new places when time allows.

[Your habits]

A slightly late riser. Commutes by subway (no car); grabs a mid-morning coffee to stay focused; cooks at home on quiet evenings.

[Now]

Tuesday, 12:40 · you are at: your office (CBD, Chaoyang).

[Today's fixed schedule]

work 10:00–18:00 (currently on a lunch break) · dentist appointment 15:00 (Chaoyang).

[Your current need state]

hunger: high (last ate at breakfast, ∼5 h ago) · energy: moderate · social: fine.

[Your cognitive map]

(current need: *food*; ⋆ = matches your current need)

Near office — Chaoyang CBD

⋆	Old Place Sushi	Japanese restaurant	visited 5×, 0.8 km
⋆	Lanzhou Beef Noodles	noodles	visited 3×, 1.2 km
	Blue Cup Coffee	café	visited 7×, 0.5 km
	Chaoyang Park	park	visited 2×, 2.6 km

Near home — northern suburbs (∼18 km)

Corner breakfast shop	breakfast spot	visited 24×, 18.2 km
Daily convenience store	convenience	visited 18×, 18.0 km
Fresh Market	grocery	visited 9×, 17.6 km

Heard-of regions (not yet visited): Sanlitun · Wangfujing · Guomao

[Episodic memory]

Pending obligation: buy groceries — shopping, due today, urgency high

Upcoming plan: dentist appointment — today 15:00, Chaoyang

Recent days: Mon — busy, worked late, ate out; Sun — relaxed, cooked at home

Reflection insight: "ate out 4× this week already — maybe cook at home more" *(importance 6)*

[Done today]

07:35–08:00 breakfast (home) ✓

Decide and output JSON:

```
{
  "next_activity": "breakfast|lunch|dinner|coffee|work|go_home|
                    shopping|leisure|social|errand|rest|...",
  "decision_type": "fixed | open",
  "location": "place name  or  home/work  or  null",
  "preferences": {
    "distance": "close | medium | far | any",
    "novelty": "familiar | new | any",
    "time_budget_minutes": <int>,
    "specific_attributes": "free-text taste/keywords, may be empty"
  },
  "target_region": "region name  or  null",
  "reasoning": "one short sentence: why this choice"
}
```

F.2 DESTINATION SELECTION PROMPT

This prompt selects one destination and a jump mode from the bounded consideration set returned by retrieval; the map's knowledge enters only as soft annotations (source, familiarity, appeal), and candidates are referenced by id so the model can only pick from what its map contains.

Stage 3: destination selection prompt (Abridged example)

System: You are a Beijing resident deciding exactly where to go and how to get there. Pick one place from the candidates below and a feasible way to reach it. Output only JSON.

[Your decision]

lunch — hungry; want something quick and open to trying a new spot near the office.

[Reflection insight]

"ate at Jinyi Pot (hotpot) too often lately — try some variety."

[Done today]

07:35–08:00 breakfast (home)

[Candidates]

([L1a] visited · [L1b] heard of · [DB] newly found)

`[L1a]` **Lanzhou Beef Noodles** `(id: B0A593)` — noodles · 0.4 km · walk ∼6 min · ⋆4.3 · visited 8×, appeal 0.90

`[L1a]` **Xibei Oat-Noodle Village** `(id: B0A83E)` — northwestern · 0.9 km · walk ∼12 min · ⋆4.5 · visited 3×, appeal 0.80

`[L1b]` **Chongqing Noodles** `(id: B0C71K)` — Noodles · 1.1 km · walk ∼14 min · ⋆4.4 · heard of, never been

`[DB]` **Grandma's Kitchen** `(id: B0D22P)` — home-style · 1.6 km · transit ∼13 min · ⋆4.6 · brand new

`[DB]` **Real Kungfu** `(id: B0E88Q)` — fast food · 0.6 km · walk ∼8 min · ⋆4.0 · brand new

[Preferences]

distance: close · novelty: new · time budget: 45 min · wants: quick, light

[Exploration]

You tend to try new places — don't always fall back to the familiar.

[Constraints]

available transport: walk, bike, transit

[Note]

next fixed event: dentist 15:00 — mind the time.

Choose a destination + transport, and output JSON:

```
{ "chosen_place_id": "B0D22P",           // Grandma's Kitchen
  "transport_mode": "transit",
  "noteworthy_alternatives": ["B0E88Q"]  // only [DB] ids
}
```

F.3 REFLECTION PROMPT

Preference-reflection prompt. This scores how well each visited place fits the persona's tastes, with visit counts deliberately withheld so that preference does not collapse onto mere habit.

Preference-reflection prompt (Abridged example)

System: You are a Beijing resident doing a weekly self-review. Output only JSON, no explanation.

[Your persona / tastes]

Female, 29, marketing specialist. A foodie who likes specialty restaurants and quiet independent cafés.

[Task]

Rate each place by *taste match*, NOT by how often you go.

Set your visit counts aside entirely (going often may be habit or convenience, not liking). Judge each place only by how well it fits your persona and tastes, giving an appeal $\in [0, 1]$:

· fits your taste / persona (foodie $\leftrightarrow$ specialty spots) $\rightarrow$ high

· purely functional (bank, grocery, chain) $\rightarrow$ low

[Places you have visited]

(visit counts deliberately hidden)

`id:B000A0B593` Old Place Sushi (restaurant)

`id:B000A83C8E` Lanzhou Beef Noodles (restaurant)

`id:B000B21X7Q` Corner Grocery (daily-use)

`id:B000C55K2P` Indie Café (café)

[Output JSON]
(place_id = the id only, no name/parens):

```
{ "appeals": [
    {"place_id": "B000A0B593", "appeal": 0.0-1.0,
     "reason": "one-line taste basis"}
] }
```

Weekly-reflection prompt. This weekly call distills the past week's statistics into three kinds of insight—pattern, self-image, and emerging need—each grounded in specific numbers, which are stored in episodic memory and reused in later decisions.

Weekly reflection prompt (Abridged example)

[System instruction]

You are a Beijing resident reflecting on your past week. Base every observation on the specific numbers below — no vague generalities. Output only JSON.

[Recent pattern data]

Activity distribution (7d)	breakfast 7 · work 5 · dinner 5 · coffee 4 · · ·
Most-visited places (7d)	Jinyi Pot 9 · office 5 · breakfast shop 6
Work intensity	5/7 work-days, avg 10.2,h/day
Mood trajectory	11-04 busy · 11-05 tired · 11-06 tired · · ·

[Reflection task]

Produce three kinds of insight, each grounded in the numbers above:

`pattern`
e.g. *"ate at Jinyi Pot 9× this week — maybe craving variety"*

`self_image`
e.g. *"worked 5 days at avg 10.2,h — intensity is high"*

`need_emergence`
e.g. *"tired 6 days straight — needs rest next week"*

[Output format]

Output JSON (each insight carries an importance 1–10):

```
{
  "insights": [
    {
      "insight_type": "pattern | self_image | need_emergence",
      "content": "...",
      "importance": 1-10
    }
  ]
}
```