GenUrban: A Generative Agent-Based Framework for Urban Analysis and Planning

Cafer Avcı[1,2]*, Changlin Huang[2], Ran Yi[2], Wanqi Tang[2], Yi He[2]

[1]Civil and Environmental Engineering, Cornell University,
310 Hollister Hall, Ithaca, NY, 14853-3801, USA
[2]Systems Engineering, Carpenter Hall, Cornell University, Ithaca, NY 14853, USA
[3]Systems Engineering, Center for Transportation, Environment, and Community Health (CTECH),
Cornell University, 313 Hollister Hall, Ithaca, NY, 14853-3801, USA
ca548@cornell.edu, ch2269@cornell.edu,ry357@cornell.edu
wt322@cornell.edu, yh2323@cornell.edu

Abstract

Urban mobility systems exhibit complex, emergent behaviour that is poorly captured by traditional top-down demand models. We introduce GenUrban, a three-stage generative agent-based framework that couples large-language-model (LLM) cognition with heterogeneous urban data to synthesise population clusters, 24-hour activity schedules and hourly origin–destination (OD) flows. The pipeline comprises (1) PopulationAgent, which partitions census microdata (ACS) into archetypal demographic clusters; (2) ActivityAgent, which maps each cluster to hour-by-hour participation fractions across five canonical activities; and (3) MobilityAgent, which transforms those fractions into OD records consistent with real-time TomTom traffic and building-footprint context (OvertureMaps). All agents are implemented as ConversableAgents in AutoGen and powered by the Grok-3-beta LLM, with a lightweight user-proxy layer providing on-the-fly validation. Using Manhattan, NY as a testbed, GenUrban automatically generated seven behaviourally distinct clusters whose size, age, income and travel-mode mixes reflect observed distributions . The resulting activity schedules reproduce diurnal peaks for work, meal and leisure periods, while OD flows from the 08:00–09:00 window align with empirical traffic patterns. An ablation study with LLaMA-3.1-8B shows pronounced quality degradation, confirming the importance of model scale for behavioural fidelity. GenUrban thus bridges cognitive realism and system-level planning, offering a scalable foundation for data-driven urban digital twins.

1 Introduction

Background and Motivation

Urban systems represent highly complex and adaptive ecosystems characterized by emergent behaviors, non-linear interactions, and interdependent processes across space and time. Traditional urban analysis tools often adopt static, linear, or siloed approaches that fail to capture the full extent of these dynamics. As cities continue to grow and evolve, the need for more responsive, behaviorally realistic modeling tools has become increasingly evident. Recent advances in agent-based modeling, generative AI, and real-time data integration present new opportunities for simulating how human behavior and infrastructure interact to shape urban life [Liu et al., 2024, Fraser et al., 2024]. Motivated

*Corresponding author: ca548@cornell.edu

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

by these developments, this work seeks to bridge the gap between cognitive realism and system-level planning through a new generative framework.

Conventional urban mobility models typically rely on top-down rules and simplified decision trees, which overlook the spontaneity, diversity, and adaptability of human behavior in dynamic environments. Factors such as weather, time of day, social context, and local events often lead individuals to modify their travel routines in ways that traditional models cannot easily capture. Moreover, while urban digital twins have gained prominence in infrastructure-level simulation, they often fail to reflect realistic citizen responses or social adaptation patterns [Zhang et al., 2024, Amirgholy and Gao, 2023]. This paper addresses these limitations by introducing a generative agent-based framework—GenUrban—that leverages cognitive architectures, environmental data, and social interactions to simulate emergent urban mobility patterns at both the individual and aggregated levels.

Research Objectives

To advance the capabilities of *aggregated* agent-based urban mobility simulation, this study addresses three core directions:

1. **Cluster-Level Behavior Modeling:** Develop methods for synthesizing representative population clusters—each defined by demographic summaries (age, income, education, etc.)—and for translating those cluster profiles into hour-by-hour activity participation fractions. This objective examines how aggregate attributes interact with temporal and contextual triggers to produce realistic, time-varying activity mixes that capture both routine patterns and anomalous conditions.
2. **Integrating Heterogeneous Urban Data Sources:** Design and implement a unified data pipeline that fuses census distributions (ACS), building footprints and land-use (OvertureMaps), and real-time OD Data (TomTom Mobility API). By aligning these diverse inputs, agents generate context-aware aggregated activity profiles that reflect observed urban dynamics [**?**Fraser et al., 2024].
3. **Evaluating Aggregated Realism and Scalability:** Establish a suite of validation metrics and benchmarks tailored to aggregated outputs—hourly activity fractions and regional flow matrices. This includes expert review, comparison against TomTom and NHTS benchmarks, and sensitivity analyses across different urban typologies (with Manhattan as a high-density test case) to assess both behavioral realism and computational scalability [Liu et al., 2022, Amirgholy et al., 2023].

2 Related Work

2.1 Generative AI in Mobility Analysis

Generative AI techniques are increasingly used to enhance mobility analysis by creating realistic synthetic populations and travel behaviors. A primary motivation is to overcome data limitations in transportation research. Mobility data contains rich information but is often limited or withheld due to privacy concerns. For example, as few as three time-stamped locations per person can re-identify individuals in transit datasets[Kapp et al., 2023], making agencies hesitant to share detailed data. Generative models offer a way to produce substitute data that mimics real patterns without exposing personal identifiers. Recent surveys report a "heterogeneous, active field of research" around generative models for synthetic mobility data, indicating broad interest in this approach. By training on whatever data are available, generative AI can learn the statistical structure of human mobility and then synthesize new data samples that have similar characteristics. This capability benefits researchers by providing abundant data for simulations or analyses when real data are sparse.

One area where generative AI has made an impact is in synthetic population creation for travel demand models. Classical methods typically select or weight survey persons to match aggregate demographics, which risks bias and cannot create persons with novel attribute combinations. In contrast, generative models can learn the joint distribution of demographic attributes and generate entirely new individual records that still reflect the real population's characteristics. This enhances diversity in the synthetic population by including plausible "sampling zeros" while avoiding impossible "structural zeros". For example, a recent study used a deep generative model to synthesize a population for an activity-based

model and showed it could generate 23.5% additional individuals with 79% validity, markedly improving population coverage. Such advances address longstanding problems of small sample bias in travel surveys. Another study replaced the survey seed in an Iterative Proportional Fitting procedure with GAN-generated persons and achieved a better match to census targets than using the raw survey data. These results demonstrate that AI-generated synthetic populations can increase the representativeness and accuracy of inputs to mobility simulations.

Generative AI is also used to create synthetic mobility traces and behavior patterns. Researchers have applied sequence-generating models to learn how people move through space and time. For instance, models based on recurrent neural networks (RNNs) and sequence-to-sequence architectures have been trained on GPS trajectory datasets to produce realistic movement sequences. Such models can capture complex spatial-temporal dependencies that are hard to encode with manual rules. GAN-based frameworks have been developed to generate human trajectories as well, sometimes combined with inverse reinforcement learning to ensure the generated paths are rational.

2.2 Population Synthesis

Population synthesis is the task of creating a complete set of individual agents for a study area, usually by expanding a small sample to match aggregate demographics. This synthetic population is a crucial input to activity-based travel models. Traditional approaches to population synthesis include synthetic reconstruction and combinatorial optimization. A common method is iterative proportional fitting (IPF), which adjusts the weights of sample records so that the marginals of synthesized data match known totals[Kotnana et al., 2022]. Variants like iterative proportional updating (IPU) handle multi-level constraints. Such techniques are robust and simple but have limitations: they replicate survey individuals and thus cannot generate persons with attribute combinations do not present in the sample. If the sample under-represents certain minorities or rare types, the synthetic population will inherit those biases. To mitigate this, researchers have explored more advanced statistical methods. For example, Bayesian networks have been used to model the joint distribution of demographic attributes and sample new individuals from it. This was shown to better capture multi-dimensional correlations than IPF for the Swiss population. Similarly, combinatorial optimization methods select and clone entire households from the sample to match multi-way targets, sometimes using metaheuristics like hill-climbing or genetic algorithms. These methods improve the consistency of household-person characteristics but can be computationally intensive and still limited by the input sample's diversity.

In recent years, deep generative models have been introduced to population synthesis, bringing notable improvements. Borysov et al. (2019) first applied a Variational Autoencoder (VAE) to synthesize populations, demonstrating the potential of deep learning to capture complex feature dependencies in survey data. More recently, Kim and Bansal (2022) proposed a GAN-VAE approach with custom regularizations to balance diversity and realism. They addressed the key issue of sampling zeros versus structural zeros: generative models can create new attribute combinations that were not in the sample but must avoid producing impossible combinations. By penalizing unlikely combinations during training, their model achieved a synthetic population with high diversity and high feasibility: for example, the VAE generated 23.5% additional "new" individuals beyond the sample, with about 79% of them being. The GAN variant generated a slightly smaller fraction of new individuals (18.3%) but with higher precision (89% valid). Crucially, both outperformed traditional IPF-based synthesis in reproducing the true population's distributions, confirming that deep generative models can synthesize more representative populations.

Beyond single jurisdictions, researchers are now considering how to make population synthesis dynamic and transferable. One emerging idea is to condition generative models on high-level context so they can generate populations for different cities or future years under demographic change. This would be a step toward scenario-based population synthesis, important for long-term transportation planning. Another extension is linking synthetic populations to activity attributes.

2.3 Activity Pattern Generation

Activity pattern generation involves modeling the daily schedules of individuals and travel episodes that each person undertakes in a day. This is central to activity-based travel demand modeling, under the principle that travel is derived from the demand to participate in activities. Traditional approaches to activity generation can be categorized into (a) rule-based models and (b) utility-based models[Erath

et al., 2012]. Rule-based models encode behavioral heuristics or decision trees often calibrated from survey data. For instance, they might enforce that "if work duration is short, probability of adding a secondary stop after work is low," based on observed patterns. Utility-based models formulate activity scheduling as an optimization problem: individuals choose the set of activities and timings that maximize their utility. Examples include the TASHA model for Toronto and CEMDAP for Dallas–Fort Worth, which use discrete choice models to simulate activity–travel patterns. These approaches have provided valuable insights and are grounded in travel behavior theory. However, they can become extremely complex as one tries to accommodate all possible activity sequences and constraints. A recent study noted that "much less research [has] exploit[ed] the advantage of deep learning for activity generation tasks" compared to other areas like mode choice or traffic forecasting. This gap is now starting to be filled by AI-driven methods.

Data-driven and AI-based approaches to activity pattern generation have rapidly progressed in the past few years. One line of work applies deep learning models to learn activity–travel sequences from large-scale travel survey datasets or location traces. For example, Phan and Vu (2021) proposed a framework that represents each person's day as one primary activity tour plus several secondary tours. They then trained neural network classifiers and regressors with embedded representations of categorical features to predict activity types and start/end times for each segment of the tours. The model captured observed patterns such as the typical timing of work and school activities with high accuracy. This showed that with enough data, a purely learning-based approach can replicate the timing distribution of activities without explicitly coding those peaks. Other studies have used sequence modeling networks – e.g. Long Short-Term Memory (LSTM) recurrent networks or Transformer-based models – to generate an entire day's activity schedule as a sequence of symbols (H = Home, W = Work, E = Education, S = Shop, etc.) with associated timestamps. Liang et al. (2022) train such models on longitudinal survey data and demonstrate that the networks can learn realistic patterns of daily behavior, including the propensity to do certain activities in sequence and the typical durations of each. Importantly, these deep models can incorporate multiple features simultaneously and capture non-linear relationships that would be very cumbersome in a traditional model. A recent "Deep Activity Model" by Huang et al. (2024) exemplifies the state-of-the-art: it uses a Transformer encoder–decoder network to generate activity sequences for individuals, conditioning on both personal attributes and household context. By encoding household members' information, it ensures that the simulated activities reflect intra-household. This model was trained on a national travel survey and successfully reproduced distributions of tour lengths, activity durations, and trip start times, while also maintaining logical consistency within households. Such deep generative models of daily schedules effectively learn the rules and trade-offs of activity-time allocation directly from data, rather than requiring manual utility functions.

An entirely new avenue is the use of Large Language Models (LLMs) as generative agents for activity planning. Wang et al. (2024) introduced an LLM-based agent framework for personal mobility generation[Wang et al., 2024]. The idea is to prompt an LLM with information about an individual and have it "imagined" a realistic day plan for that persona in natural language, which is then parsed into a structured itinerary. The authors address key challenges like aligning the LLM's output with real travel survey data to avoid unrealistic plans and ensuring consistency in the agent's story. They develop a self-consistency approach where the LLM generates multiple candidates plans and checks them against known patterns, and a retrieval-augmented strategy that feeds the LLM contextual facts to ground its outputs. The result is a system that can produce highly diverse activity patterns that are also interpretable and statistically like observed behavior. This is a notable innovation as it brings together the world knowledge and reasoning ability of LLMs with the quantitative rigor of mobility models. While still experimental, early results show the LLM-based approach handling semantic nuances that hard-coded models might miss. It also opens the door to explainable simulation, as the agents can describe why they made certain choices.

2.4 Agent Interaction Frameworks

A distinguishing feature of generative agent-based models is the potential for agent–agent interactions within the simulation. In conventional travel modeling, individual travelers' decisions are usually treated as independent, aside from shared constraints[Hackney and Marchal, 2008]. Realistically, however, people's travel choices can be influenced by others: family members coordinate schedules, friends decide to meet up, information or social pressure can spread through communities affecting mode choices, etc. Capturing these peer-to-peer influences is challenging but can greatly enrich urban

mobility analysis[Arentze and Timmermans, 2008]. Early activity-based models did account for some interactions. Yet, interactions beyond the household were largely ignored due to lack of data and the complexity of modeling social influence. As a result, traditional models might systematically mis-predict certain behaviors. For example, assuming travelers act in isolation might fail to predict the synchronization of travel seen during events or the diffusion of new behaviors through social networks.

Recent research is beginning to incorporate explicit agent interaction frameworks in mobility simulations. One approach is to integrate social network models with travel demand models. For instance, a multi-agent simulation by Zhu et al. created a synthetic social network among agents and allowed information about a new transit service to spread via person-to-person communication; the result was a more realistic uptake curve for the service compared to assuming all travelers learn about it independently. Similarly, Hackney and Axhausen (2006) linked a social network model with an activity generation model, showing that the formation of social ties and activity-travel patterns can co-evolve – people who frequently interact often align their activities or travel together. Their framework allowed agents to socialize within the travel simulation, demonstrating the feasibility of simulating joint activities beyond households. The outcomes illustrated that including social interactions can change travel demand results: for example, an agent might make an extra trip they wouldn't have made alone, because a friend invited them. Such findings underscore that ignoring interactions could miss emergent phenomena. Indeed, a coupled social-travel simulation by Arentze and Timmermans (2008) found that accounting for friendship networks explained certain discretionary travel that pure economic models did not[Sunitiyoso et al., 2010]. These pioneering studies, while simplified, verify that multi-agent systems can capture peer influence in travel behavior.

Building on these ideas, generative agents in urban mobility projects are beginning to incorporate communication protocols that enable agents to interact when certain conditions are met. A key example is designing agents to recognize shared geolocation and time-frame overlaps. For instance, if multiple agents are at the same place at the same time (e.g. coworkers in the office at 5 PM), the simulation can trigger an interaction – perhaps they decide to share a ride home or go to an event together. Implementing this requires a framework where agents can detect nearby agents and send/receive messages. In multi-agent systems research, standardized messaging languages exist, but in transport contexts simpler rule-based triggers may suffice. The literature has a few examples: in a shared mobility study, Martinez et al. (2017) let ride-hailing requests from agents be matched in real-time, effectively allowing agents to "team up" for shared rides – a form of interaction through a mediator algorithm. Another example is in traffic signal control: multiple studies have used multi-agent reinforcement learning where intersection agents communicate their congestion levels to neighbors to coordinate green lights[Zhao et al., 2020]. While this is infrastructure-focused, it showcases how communication can improve system performance. By analogy, if traveler agents could communicate, the collective outcome might be more efficient equilibrations than if each agent acts purely individually.

Generative agents are well-suited to exploring such interactive dynamics because their behaviors are not scripted and can adapt. Researchers envision scenarios like agents sharing information about parking availability via a decentralized network or simulating the peer pressure effect of co-workers: if a critical mass of agents start taking transit, others in their social circle may follow, altering the mode split outcome. To enable this, one needs both a realistic social network and a protocol for interaction. Some studies create synthetic social networks by spatial-temporal overlap. Others import known networks when available. The communication protocols can be direct messaging or indirect influence. Park et al.'s generative agents in a sandbox game even initiated conversations when meeting[Park et al., 2023]; translating this to urban mobility, one could imagine agents conversing when carpooling or when waiting at transit stops, affecting their satisfaction or future choices. While such richness is mostly conceptual at this stage, initial frameworks have been proposed. For example, Schirmer et al. (2022) outline a model where agents have "opinion states" about modes that can change if they discuss with others; this model, when simulated, produced clusters of mode adoption like real-world observations of neighborhood effect in mode choice.

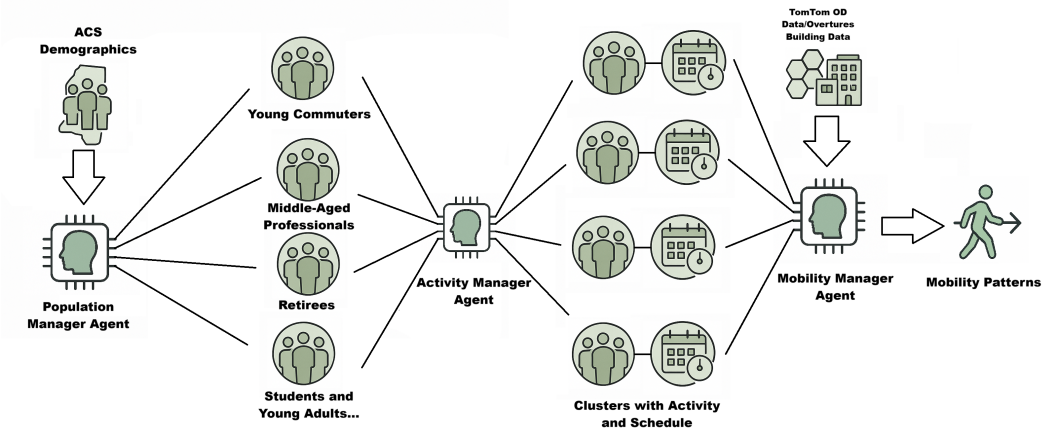

Figure 1: The System Design and Data Flow of the GenUrban Framework

3 Methodology

3.1 Data Input

Our framework relies on three core data streams: demographics distribution, building footprints, and Real-time OD pair over regions. We retrieve age, gender, educational attainment, and income class breakdowns for any U.S. county via the American Community Survey (ACS) API. A single function call returns the latest population counts per strata. We use the overturemaps Python package to load building geometries, categories (e.g., residential, commercial, recreational), and geocoordinates. This inventory supplies valid building id values for all individual and aggregated activity assignments. Real-time OD records between customized hexagons are pulled from the TomTom Mobility API, provide insights for the estimation of generation of the mobility patterns for each population cluster. (Note that, I am not sure about here whether we should talk in depth about the hexagon, need more information to fill this part.)

3.2 System Design

As shown in Fig. 1, our aggregated pipeline features three sequential stages, each executed by a specialized LLM-powered "manager" agent. The PopulationAgent ingests county-level ACS distributions and partitions the population into a small number (e.g., 5-10) of archetypal clusters—each defined by cluster-id, total count, and a compact summary of attributes distribution such as age, income, education attainment, and so on. The ActivityAgent takes each cluster and a fixed list of activity types (sleep, work, meal, errand, leisure). It divides the 24-hour day into hourly bins and, for each bin and activity, estimates the fraction of the cluster engaged. These fractions sum to one across all activities in each hour. Finally, the MobilityAgent transforms the hour-by-hour activity fractions into origin–destination records, while simulating the real worlds OD records from TOMTOM.

All of our agents are implemented as 'ConversableAgent' instances in the AutoGen framework, each wrapping the grok-3-beta model served from x.ai, and are driven by dedicated 'ChatSession' objects with attached 'MemoryBuffer' histories. We also interpose a custom UserProxy layer on each session that intercepts every manager's request and response to log, validate, and, if necessary, correct outputs on the fly—ensuring full traceability and quality control throughout population clustering, activity aggregation, and flow-matrix generation.

(Note, could add a section of comparison with the lighter llama3.1 8b, but it could led the paper exceed the page limit.)

4 Results

Manhattan (New York County, New York) serves as our experimental testbed. As an insular, high-density urban environment with a resident population of 1,627,788, Manhattan features a complex built landscape—encompassing residential, commercial, religious, educational, and industrial structures—and a population exhibiting pronounced heterogeneity in age, income, race, and educational attainment. These characteristics make Manhattan an ideal setting for rigorously testing and validating our aggregated agent-based mobility simulation framework.In the following subsections, we will exihibit our results at each stage of the simulation.

4.1 Population Clusters

As shown in Tables 1, 2 and 3, the PopulationAgent autonomously generated seven archetypal clusters—each endowed with a distinct identity and attribute profile—within the New York County testbed. Importantly, no explicit instructions were provided regarding either the criteria for differentiation or the exact number of clusters; the AI agent determined both entirely through its internal clustering logic. As illustrated in Figure 2, the PopulationAgent successfully discriminated among these clusters, which exhibit unique demographic compositions and substantially varied population sizes and characteristics—factors that are expected to exert a pronounced influence on simulated mobility patterns. The mapping of cluster name to cluster id is presented in Table 4

Cluster ID	Population	Age	Income
YC_01	250,000	(18–34: 70%, 35–54: 20%, 55+: 10%)	(Low: 30%, Middle: 50%, High: 20%)
MP_02	300,000	(18–34: 15%, 35–54: 70%, 55+: 15%)	(Low: 10%, Middle: 30%, High: 60%)
RT_03	200,000	(18–34: 5%, 35–54: 15%, 55+: 80%)	(Low: 40%, Middle: 40%, High: 20%)
LI_04	250,000	(18–34: 40%, 35–54: 40%, 55+: 20%)	(Low: 70%, Middle: 25%, High: 5%)
FH_05	350,000	(18–34: 20%, 35–54: 60%, 55+: 20%)	(Low: 20%, Middle: 50%, High: 30%)
SY_06	150,000	(18–34: 85%, 35–54: 10%, 55+: 5%)	(Low: 60%, Middle: 30%, High: 10%)
HE_07	127,788	(18–34: 10%, 35–54: 60%, 55+: 30%)	(Low: 5%, Middle: 15%, High: 80%)

Table 1: Cluster IDs with Population, Age, and Income

Cluster ID	Education	Gender	Employment
YC_01	(HS: 25%, SC: 35%, BA+: 40%)	(M: 52%, F: 48%)	(FT: 80%, PT: 15%, U: 5%)
MP_02	(HS: 10%, SC: 20%, BA+: 70%)	(M: 50%, F: 50%)	(FT: 85%, PT: 10%, U: 5%)
RT_03	(HS: 40%, SC: 30%, BA+: 30%)	(M: 45%, F: 55%)	(Retired: 90%, PT: 10%)
LI_04	(HS: 50%, SC: 30%, BA+: 20%)	(M: 55%, F: 45%)	(FT: 60%, PT: 30%, U: 10%)
FH_05	(HS: 30%, SC: 30%, BA+: 40%)	(M: 48%, F: 52%)	(FT: 70%, PT: 20%, U: 10%)
SY_06	(HS: 20%, SC: 50%, BA+: 30%)	(M: 50%, F: 50%)	(PT: 50%, U: 30%, FT: 20%)
HE_07	(HS: 5%, SC: 10%, BA+: 85%)	(M: 55%, F: 45%)	(FT: 90%, PT: 5%, U: 5%)

Table 2: Cluster IDs with Education, Gender, and Employment

Cluster ID	Household	Travel Behavior
YC_01	–	(Transit: High, Car: Moderate)
MP_02	–	(Car: High, Transit: Moderate)
RT_03	–	(Car: Moderate, Transit: Low)
LI_04	–	(Transit: High, Car: Low)
FH_05	(With Children: 80%)	(Car: High, Transit: Low)
SY_06	–	(Transit: High, Walk/Bike: High, Car: Low)
HE_07	–	(Car: High, Transit: Moderate)

Table 3: Cluster IDs with Household and Travel Behavior

Cluster ID	Cluster Name
YC_01	Young Commuters
MP_02	Middle-Aged Professionals
RT_03	Retirees
LI_04	Low-Income Workers
FH_05	Family Households
SY_06	Students and Young Adults
HE_07	High-Income Executives

Table 4: Mapping of Cluster IDs to Cluster Names

Figure 2: Population Cluster Attributes Distribution Radar Diagram

4.2 Activity and Schedule Patterns

As illustrated in Figure 3, the generated activity schedules exhibit empirically plausible patterns. Sleep periods are predominantly allocated to nighttime hours, while work activities occur chiefly during the daytime. Meal times cluster around 06:00–08:00, 11:00–13:00, and 17:00–19:00, and leisure activities peak in the evening. Moreover, clear inter-cluster distinctions emerge: for instance, the Retirees cluster engages in no work activities and begins its primary sleep episode approximately two hours earlier than other groups, whereas the Students and Young Adults cluster shows reduced participation in work or study activities, reflecting the substantial proportion of individuals below school-age.

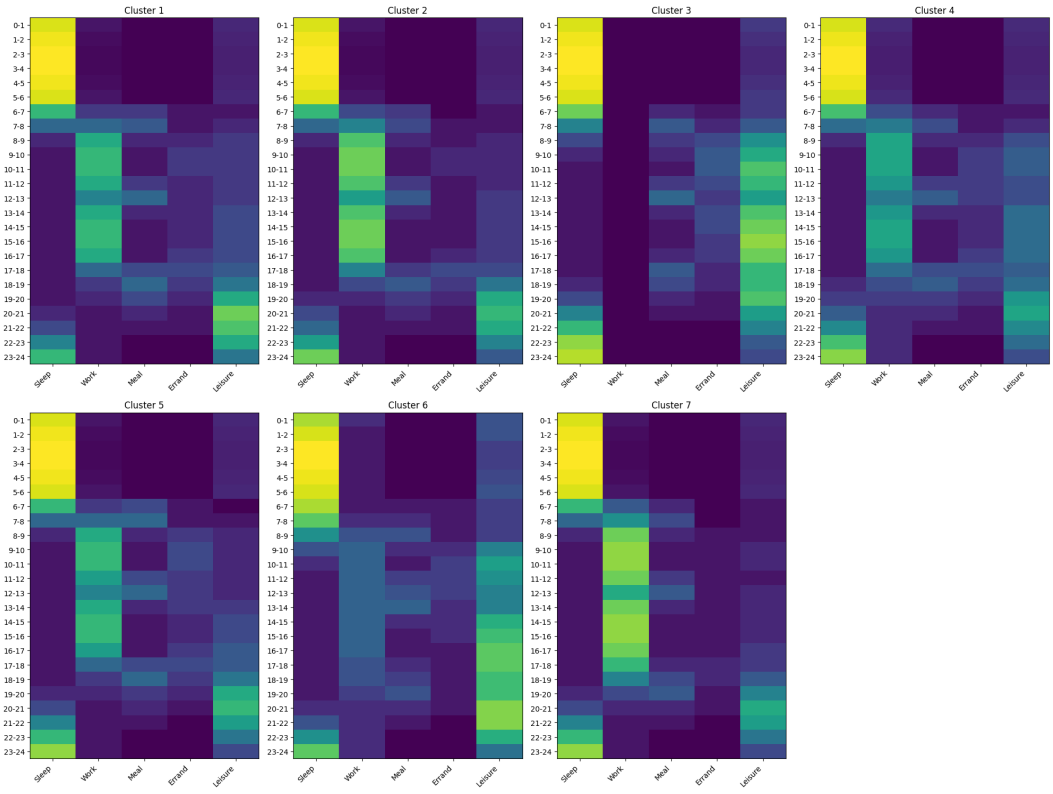

Figure 3: The Activity-Schedule Patterns of 7 Population Clusters

4.3 Mobility Patterns

Although the initial two agents achieved satisfactory results on simpler tasks, their performance degrades markedly when confronted with increased input complexity or when generating outputs that require more nuanced reasoning. In such cases, the agents frequently fail to execute the prompt as intended or outright refuse to complete the task, even when token-limit constraints are not reached [Touvron et al., 2023, Dai et al., 2019]. In contrast, our MobilityAgent successfully produces high-quality mobility data for the 08:00–09:00 interval, generating origin–destination flows from hexagon 0 to all other hexagons, as illustrated in Figure 4.

5 Discussion

To further investigate this limitation, we evaluated our framework using the lightweight LLaMA-3.1 8B model while keeping all other experimental conditions fixed. The resulting activity schedules, shown in Figure 5, exhibit substantially lower fidelity compared to those produced by Grok-3-beta. This performance degradation is consistent with established scaling laws linking model capacity to output quality [Kaplan et al., 2020]. These results confirm that the effectiveness of our framework is intrinsically tied to the scale and architecture of the underlying LLM. We therefore anticipate that incorporating future, more powerful models will yield correspondingly higher-quality synthetic mobility data.

6 Conclusion

This study presents GenUrban, a modular, LLM-driven framework that fuses demographic, land-use and traffic data to generate behaviourally plausible, aggregated urban mobility outputs. Experiments in Manhattan demonstrate that (i) LLM-powered clustering can uncover meaningful population

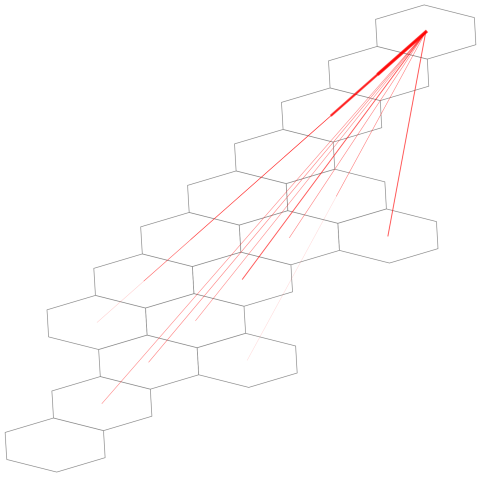

Figure 4: Manhattan (New York County) OD Flow from Hexagon 0

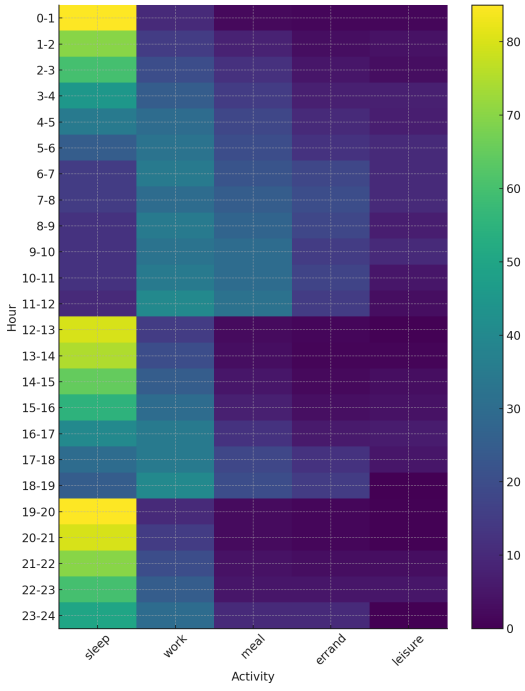

Figure 5: The Activity-Schedule Patterns Generated by Llama3:8b

archetypes, (ii) transformer-based agents reliably translate cluster attributes into temporally resolved activity mixes, and (iii) the resulting OD matrices capture spatial-temporal flow structures observed in high-density cities. Comparison with a smaller 8-billion-parameter LLaMA variant underscores a clear scaling–quality trade-off, suggesting that continued advances in foundation-model capacity will directly enhance simulation realism.

Nevertheless, GenUrban currently abstracts away explicit agent–agent interactions and relies on limited external validation. Future work will embed social-network communication protocols, incorporate dynamic feedback from real-time sensors and extend testing to diverse urban typologies. Integrating rigorous statistical goodness-of-fit metrics and stakeholder co-design loops will further strengthen the framework's policy relevance. By uniting generative AI with urban-systems science, GenUrban offers a promising pathway toward responsive, explainable and city-scale digital twins capable of informing next-generation planning and resilience strategies.

References

M. Amirgholy and H. O. Gao. Optimal traffic operation for maximum energy efficiency in signal-free urban networks: A macroscopic analytical approach. *Applied Energy*, 329:120128, 2023. doi: 10.1016/j.apenergy.2022.120128.

M. Amirgholy, M. Nourinejad, and H. O. Gao. Balancing the efficiency and robustness of traffic operations in signal-free networks. *Transportation Research Interdisciplinary Perspectives*, 19: 100821, 2023. doi: 10.1016/j.trip.2023.100821.

Theo Arentze and Harry Timmermans. Social networks, social interactions, and activity-travel behavior: A framework for microsimulation. *Environment and Planning B: Planning and Design*, 35(6):1012–1027, 2008. doi: 10.1068/b3319t.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

Alexander Erath, Pieter Jacobus Fourie, Michael A.B. van Eggermond, Sergio Arturo Ordonez Medina, Artem Chakirov, and Kay W. Axhausen. Large-scale agent-based transport demand model for singapore. Working Paper 790, Future Cities Laboratory (FCL), 2012. URL `https://www.research-collection.ethz.ch/handle/20.500.11850/306926`.

Timothy P. Fraser, Y. Guo, and H. Oliver Gao. Making moves move: Fast emissions estimates for repeated transportation policy scenario analyses. *Environmental Modelling and Software*, 178: 106084, 2024. doi: 10.1016/j.envsoft.2024.106084.

Jeremy K. Hackney and Fabrice Marchal. A model for coupling multi-agent social interactions and traffic simulation. *Arbeitsberichte Verkehrs- und Raumplanung*, 516:1–20, 2008. doi: 10.3929/ethz-a-005652382.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Alexandra Kapp, Julia Hansmeyer, and Helena Mihaljević. Generative models for synthetic urban mobility data: A systematic literature review. *arXiv preprint arXiv:2407.09198*, 2023. doi: 10.48550/arXiv.2407.09198. URL `https://arxiv.org/abs/2407.09198`.

Srihan Kotnana, David Han, Taylor Anderson, Andreas Zufle, and Hamdi Kavak. Using generative adversarial networks to assist synthetic population creation for simulations. In *Proceedings of the 2022 Annual Modeling and Simulation Conference (ANNSIM)*, pages 1–12. IEEE, 2022. doi: 10.23919/ANNSIM55834.2022.9859422.

Y. S. Liu, M. Tayarani, and H. O. Gao. An activity-based travel and charging behavior model for simulating battery electric vehicle charging demand. *Energy*, 258:124938, 2022. doi: 10.1016/j.energy.2022.124938.

Y. S. Liu, M. Tayarani, F. You, and H. O. Gao. Bayesian optimization for battery electric vehicle charging station placement by agent-based demand simulation. *Applied Energy*, 375:123975, 2024. doi: 10.1016/j.apenergy.2023.123975.

Joon Sung Park, Joseph C. O'Brien, Carrie J. 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 (UIST '23)*, pages 104–117. Association for Computing Machinery, 2023. doi: 10.1145/3586183.3606763.

Yos Sunitiyoso, Erel Avineri, and Kiron Chatterjee. Complexity and travel behaviour: Modelling influence of social interactions on travellers' behaviour using a multi-agent simulation. In Elisabete A. Silva and Gert De Roo, editors, *A Planner's Encounter with Complexity*, pages 241–266. Ashgate, Aldershot, UK, 2010. ISBN 978-1-4094-0265-7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aman Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Jiawei Wang, Renhe Jiang, Chuang Yang, Zengqing Wu, Makoto Onizuka, Ryosuke Shibasaki, Noboru Koshizuka, and Chuan Xiao. Large language models as urban residents: An llm agent framework for personal mobility generation. *Advances in Neural Information Processing Systems*, 37:124547–124574, 2024. doi: 10.48550/arXiv.2402.14744.

Q. Zhang, Y. S. Liu, H. O. Gao, and F. You. A data-aided robust approach for bottleneck identification in power transmission grids for achieving transportation electrification ambition: A case study in new york state. *Advances in Applied Energy*, 14:100173, 2024. doi: 10.1016/j.adapen.2024.100173.

Y. Zhao, G. Xu, Y. Duy, and M. Fang. Learning multi-agent communication with policy fingerprints for adaptive traffic signal control. In *Proceedings of the IEEE 16th International Conference on Automation Science and Engineering (CASE)*, pages 1061–1066. IEEE, 2020. doi: 10.1109/CASE48305.2020.9216835.