Towards Efficient and Evidence-grounded Mobility Prediction with LLM-Driven Agents

Anonymous ACL submission

Abstract

Individual-level mobility prediction is essential for urban simulation, transportation planning, and policy-making. While supervised deep learning methods achieve strong accuracy, they require task-specific training and produce opaque predictions. Recent LLM-based approaches offer interpretability but suffer from high token cost, sensitivity to prompt design, and an inability to dynamically seek additional evidence when faced with ambiguous cases. We propose AgentMob, a training-free LLM-driven agent framework that formulates next-location prediction as an evidence-grounded decision process. The agent operates a toolbox of four mobility-analysis tools and adaptively controls its reasoning depth: when historical regularity is strong, it resolves predictions in a single fast-path step; when signals conflict, it iteratively gathers and cross-checks statistical evidence before committing. Experiments on two GPS-based mobility datasets show that AgentMob outperforms all existing LLM-based methods, improving Acc@1 by up to 1.92 percentage points and reducing spatial error to 2.51 km, and narrows the gap with supervised baselines to within 6.74 points in the same time. All results are achieved without task-specific fine-tuning. Ablation and case studies further reveal that the framework's effectiveness depends critically on the backbone model's tool-use competence, and that the adaptive evidence-gathering mechanism is key to resolving ambiguous predictions that defeat prompt-engineering approaches.

1 Introduction

Accurate prediction of human mobility is fundamental for informed decision-making in urban planning (Zheng et al., 2014), transportation management (Luca et al., 2021), and public health (Barbosa et al., 2020). Individual mobility prediction serves as a cornerstone for many valuable services like location-based services (Zheng, 2015), and targeted public health intervention (Oliver et al., 2020).

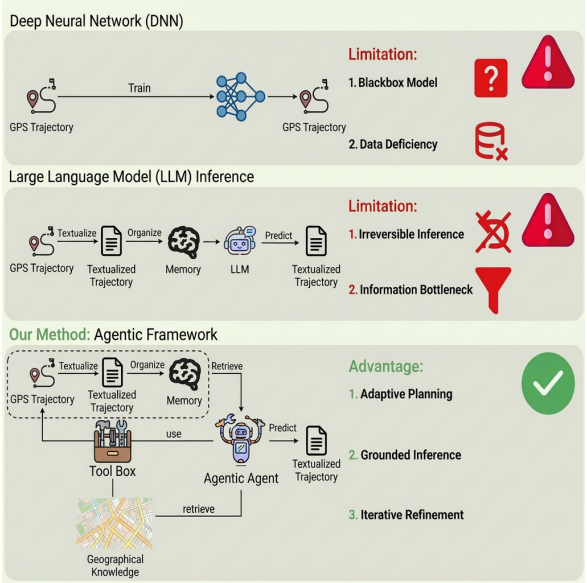

Figure 1: Comparison of the proposed agentic framework with existing approaches.

Various sequence-based deep learning models, including RNNs (Feng et al., 2018), Transformers (Vaswani et al., 2017), and their variants have achieved strong predictive accuracy, yet remain computationally expensive and produce opaque outputs with no insight into *why* a particular location was predicted. Recent efforts have turned to Large Language Models (LLMs) (Zhao et al., 2026; Zhang et al., 2024) for mobility prediction. Fine-tuning approaches (Feng et al., 2024; Li et al., 2024a) inherit the high training cost, while prompt-based methods (Wang et al., 2024b; Calderón et al., 2025) avoid retraining and compress all evidence into a single

static prompt. Agent-like frameworks (Feng et al., 2025; Li et al., 2025) further introduce structured workflows yet still follow rigid, pre-scripted reasoning pipelines. Despite this progress, the prevailing paradigm remains limited by several intractable bottlenecks.

Uninterpretability and Data Deficiency. Training sequence-based models demands large-scale trajectory data and substantial computational resources; retraining from scratch for each new city further amplifies the cost. Moreover, supervised models are tightly bound to their training distribution and provide no explicit evidence chain, leaving predictions uninterpretable and brittle in out-of-distribution scenarios.

Irreversible Inference and Information Bottleneck. Current LLM-based approaches follow a fixed, single-pass pipeline: the model reads the context, generates a prediction, and commits to it. There is no mechanism to revisit or refine the reasoning once it is produced. This problem is worsened by the limited context window, which forces the model to compress rich spatiotemporal histories into a bounded prompt. Important signals may be lost and fine-grained details quietly discarded in this process. As a result, prediction quality depends entirely on what the model captures in one forward pass, with no structured way to verify or correct the output.

To address these bottlenecks, we propose AgentMob, an LLM-based tool-integrated agent that formulates mobility prediction as an *efficient, evidence-grounded* decision process with autonomous tool-invocation capabilities. As illustrated in Figure 1, our work deviates from existing paradigms through several key innovations. We validate the framework with both open-source and closed-source LLMs on two GPS-based trajectory datasets. Experimental results show that AgentMob achieves competitive performance with supervised baselines and substantially outperforms all existing LLM-based approaches.

In summary, our contributions are as follows:

- **Agentic workflow for mobility prediction**: We are the first to recast next-location prediction as an evidence-grounded decision-making task for LLM agents, moving beyond prompt-only inference and rigid scripted pipelines.
- **Efficient reasoning with evidence-grounded tool invocation**: We develop a tool-augmented mobility agent that combines context retrieval, with adaptive evidence gathering, which includes behavioral analysis and geographic information to support interpretable reasoning over GPS-derived spatial regions.
- **Training-free Mobility prediction with Strong Performance:** We introduce an adaptive tool-calling mechanism that allocates reasoning effort based on prediction difficulty, and show its effectiveness across both closed- and open-source LLMs on two mobility benchmarks.

2 Related Work

2.1 Individual Next Location Prediction

Next location prediction has evolved from Markov chains (Rendle et al., 2010; Gambs et al., 2012) to deep models based on RNNs (Feng et al., 2018; Liu et al., 2016), Transformers (Lian et al., 2020; Luo et al., 2021; Qin et al., 2022; Sun et al., 2024), and GNNs (Wang et al., 2024c; Wu et al., 2024), aiming to capture complex spatiotemporal dependencies via training on large-scale mobility data. Recently, LLM-based approaches(Wang et al., 2023, 2024a; Liang et al., 2024) have framed mobility as textual generation or planning tasks. Early studies (Li et al., 2024a,b) fine-tune LLMs for mobility generation as a text generation task, while other frameworks(Feng et al., 2025; Ju et al., 2025; Zhong et al., 2025; Liu et al., 2025) further integrate memory and urban knowledge for high-level prediction and simulation. However, these systems largely rely on rigid, manually scripted workflows that lack the procedural autonomy and grounded tool-invocation capabilities characteristic of genuine agents.

2.2 LLM-driven Agent

LLM-driven agents represent an extension of Large Language Models into autonomous entities capable of reasoning and interaction.

Building on foundational instruction-following (Ouyang et al., 2022), research has progressed toward multi-agent systems that emphasize autonomous cooperation (Li et al., 2023; Wu et al., 2023) and modular scalability (Hu et al., 2025; Tang et al., 2025). However, their reliance on generalized conversational patterns often fails to capture the nuances of human action forecasting. Another line of research utilizes agents for social simulation (Park et al., 2023) and consistent data generation (Sengupta et al., 2024), yet these approaches remain primarily retrospective, focusing on past trajectories rather than the anticipatory reasoning required for future states. To expand capabilities, recent works integrate external tools (Nakano et al., 2021; Qin et al., 2023), shifting agents from static reasoning to interactive problem-solving. Nevertheless, current tool-use frameworks are predominantly reactive, lack the proactive architecture needed for predicting future states under uncertainty.

3 Methodology

3.1 Pipeline of LLM-based Tool Agent for Next Location Prediction

As illustrated in Figure 2, we present a tool-integrated agent workflow for mobility prediction based on smolagents (Roucher et al., 2025), where an LLM uses a set of mobility-analysis tools through an adaptive, multi-step reasoning process. Typical reasoning processes are provided in Appendix. A key advantage of this workflow is its flexibility, allowing the agent to dynamically adjust the breadth and depth of its evidence gathering. It puts more reasoning effort on ambiguous cases while resolving straightforward ones efficiently. As automatically collected raw GPS data are primarily in numerical form, we first discretize coordinates into spatial units—either administrative polygons or uniform grid cells, depending on the dataset—each paired with a short textual description based on their urban function, so that the LLM can reason about locations semantically (see Section 4.1.1 for details). The prediction task then is left to let the LLM select, from among these spatial units, the one the user will visit next.

Before accessing external tools, the agent first performs a fast-path check based on historical regularity. If the historical data indicates a strong historical regularity for the current weekday and hour in the data, the agent directly returns that location as the prediction. On the other hand, when a strong regularity is absent, the framework proceeds to the tool-calling reasoning process.

In tool-calling mode, the agent gathers evidence by invoking mobility-analysis tools from the toolbox (Section 3.2). The toolbox contains tools for two complementary purposes: (1) historical behavioral patterns and memories of mobility under similar conditions; (2) geographical information. The agent decides autonomously whether additional evidence is required, iteratively invoking tools until sufficient confidence is reached; the loop is capped at ten iterations.

Finally, after obtaining enough confidence on all collected evidence, the agent predicts the next location.

3.2 Mobility Analysis Toolbox

As introduced in Section 3.1, the agent interacts with a set of mobility analysis tools to gather behavioral memory and geographical information for prediction. These tools expose structured statistical information derived from historical trajectories and can be invoked by the agent during the reasoning process.

Mobility Context Retriever Given a target timestamp, this tool summarizes the user's recent context mobility behavior. It returns his or her recently visited locations sequential over the past several hours and historical visitation statistics for the same weekday and hour. These signals provide both a real-time snapshot of the user's recent movement and a temporal reference based on historical mobility patterns.

Geographical Information Retriever This tool provides geographical information associated with locations, so that it enables the LLM to perceive the functional characteristics of urban areas and spatial scale. It returns geographical distance from current location to candidate locations as well as a textual description of the candidate locations. These signals allow the agent to account for spatial constraints that are absent from behavioral history alone, ensuring predictions

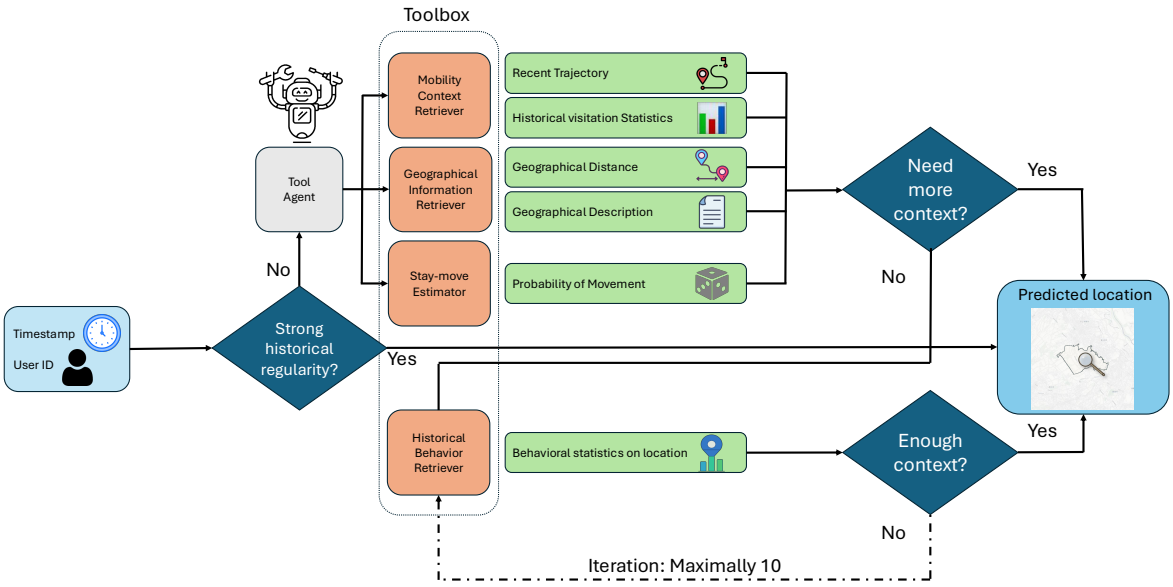

Figure 2: Overview of the AgentMob pipeline: efficient workflow, adaptive tool integration, and evidence-grounded reasoning.

remain geographically plausible rather than driven purely by pattern-level correlations.

Stay-Move Estimator Given the current location and target timestamp, this tool evaluates the statistical probability of remaining at the current location or leaving it. It compares the user's current stay duration with historical stay statistics for the same location and returns an estimated probability of movement derived from historical mobility data.

Historical Behavior Retriever If the agent is not confident on its judgement in the initial step, this tool will be adopted to retrieve behavioral statistics from historical records. Specifically, it identifies past visits to the same or nearby locations under comparable temporal contexts (e.g., similar time of day and day of week) and summarizes the user's recurring mobility patterns at those locations — such as typical dwell times, frequent next destinations, and visit regularity. These empirical priors serve as supplementary evidence to disambiguate uncertain predictions.

4 Experiment

4.1 Experimental Setup

4.1.1 Datasets

We evaluate our method on two datasets with different spatial granularities, both collected automatically via GPS sensors in mobile phones at fixed time granularity. Unlike the check-in data used in most prior work, these GPS traces record complete daily movements. We convert continuous GPS data into Japanese third-level polygons with textual descriptions of their urban function generated from available metadata (POI, region name), and select the 100 most active users from each dataset (Table 1).

BW (Blogwatcher, Inc., 2024): Originally GPS, and converted into third-level administrative polygons in the Tokyo metropolitan area, annotated with hierarchical region names and description of function.

YJMob100K (Yabe et al., 2024): originally discretized into anonymized 500 m×500 m grid cells without real region names, but provied with the information of POIs; we generate textual descriptions via LLM-based POI summarization (Appendix A).

4.1.2 Baselines

We compare against two supervised models and four LLM-based methods. **DeepMove** (Feng et al., 2018): an attentional RNN that captures periodic patterns and long-term user preferences. **Transformer** (Vaswani et al., 2017): a vanilla Transformer. **AgentMove** (Feng et al., 2025) decomposes zero-

Dataset	YJMob100K	BW
Records	310,546	427,248
Geo. Location	Anonymized	Tokyo Area
Visited Locations	12,430 cells	4,188 polygons
Move Transitions	74.7%	28.6%
Stay Transitions	25.3%	71.4%
Avg. unique locations	413.7	104.8

Table 1: Dataset statistics for BW and YJMob100K.

shot prediction into spatial-temporal memory, world knowledge, and collective pattern modules, integrated via LLM reasoning. **LLM-Mob** (Wang et al., 2023) reformulates prediction as in-context learning with structured trajectory prompts encoding historical and context stays. **TrajLLM** (Ju et al., 2025) converts mobility sequences into textual representations and relies on LLM sequential reasoning for prediction. **LLM Urban Residents** (Wang et al., 2024a) models individuals as LLM agents that generate trajectories conditioned on identified activity patterns and retrieved daily motivations. We select Qwen3-8b (Yang et al., 2025) and GPT4.1-mini (OpenAI, 2025) LLM backbones in all LLM

4.1.3 Evaluation Metrics

For each test instance i, the model outputs a ranked prediction list, and the ground-truth next location is ℓ_i. We report three metrics:

- **Acc@1.** This metric indicates whether the top-1 prediction exactly matches the ground truth:

$$\text{Acc@1} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\hat{\ell}_{i,1} = \ell_i],$$

  where N is the number of test instances and $\mathbb{I}[\cdot]$ is the indicator function.

- **MRR@5.** Mean Reciprocal Rank at 5 evaluates the rank position of the ground truth within the top-5 predictions. Let r_i be the rank of ℓ_i in the top-5 list ($r_i = \infty$ if absent):

$$\text{MRR@5} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{I}[r_i \le 5]}{r_i}.$$

- **Mean Distance (Dist.).** Mean Distance measures the average geographic distance between the top-1 prediction and the ground truth:

$$\text{Dist.} = \frac{1}{N}\sum_{i=1}^{N} d(\hat{\ell}_{i,1}, \ell_i),$$

  where $d(\cdot,\cdot)$ denotes the Haversine distance (km).

4.2 Main Results

4.2.1 Performance Against Baselines

Table 2 presents the comparison results with baseline methods on the BW and YJMob100K datasets.

Overall, our method achieves the best performance among all LLM-based training-free approaches and remains competitive with supervised baselines (underlined metrics means close performance). On the BW dataset with GPT-4.1-mini, it attains 66.30% Acc@1 and 76.33% MRR@5 that surpasses the strongest LLM-based baseline (LLM Urban Residents) by 1.92% and 3.44%, respectively. While high Acc@1 and MRR@5 confirm that the framework correctly predicts which location a user will visit, they do not guarantee that predictions are spatially sound. A method can achieve high selection accuracy by memorizing frequent destinations without understanding the underlying spatial patterns. The mean distance metric addresses this gap. At 2.51 km, the lowest among all LLM-based methods, it shows that our predictions stay close to the ground truth in geographic space, even when they do not hit the exact location.

Results on YJMob100K follow a similar trend: our method leads all LLM-based baselines with 31.81% Acc@1 and 45.97% MRR@5, while maintaining the smallest average distance (4.25 km).

Among all supervised baselines, the Transformer achieves the highest Acc@1 on both datasets, benefiting from its strong sequential modeling capacity through end-to-end optimization. Notably, our training-free framework reaches comparable performance without any task-specific fine-tuning, demonstrating that structured tool-augmented reasoning can recover much of the predictive power typically obtained through supervised learning—while additionally providing full decision interpretability. Taken together, these results show that the proposed approach generalizes across datasets with different spatial scales and mobility patterns.

Among the LLM-based baselines, LLM Urban Residents (Wang et al., 2024a) performs

Table 2: Performance comparison of mobility prediction methods on BW and YJMob100K. **Bolded** numbers indicate the best performance among all models, while underlined numbers represent the best results among LLM-based methods.

Method	Backbone	BW			YJMOB100K		
		Acc@1	MRR@5	Dist.	Acc@1	MRR@5	Dist.
Deep Neural Network	DeepMove	72.13%	78.92%	2.21	31.52%	44.86%	4.86
	Transformer	**73.04%**	**79.67%**	**2.18**	**33.12%**	**46.12%**	4.81
AgentMove	GPT-4.1-mini	58.52%	70.16%	3.53	29.25%	41.33%	4.77
	Qwen3-8b	52.58%	64.09%	4.21	29.41%	41.84%	5.36
LLM-Mob	GPT-4.1-mini	55.08%	66.10%	3.52	21.38%	35.42%	5.98
	Qwen3-8b	56.19%	65.53%	3.45	23.57%	36.50%	5.84
TrajLLM	GPT-4.1-mini	53.08%	64.33%	6.58	24.25%	37.21%	11.57
	Qwen3-8b	52.58%	64.04%	4.71	28.12%	42.23%	5.16
LLM Urban Res.	GPT-4.1-mini	64.38%	72.89%	2.59	26.59%	40.02%	5.21
	Qwen3-8b	58.93%	68.00%	3.29	25.89%	39.51%	5.90
Ours	GPT-4.1-mini	66.30%	76.33%	2.51	31.81%	45.97%	**4.25**
	Qwen3-8b	62.65%	74.30%	2.90	30.56%	45.27%	**4.39**

best on BW, likely because its motivation-driven activity modeling aligns well with the higher temporal regularity in that dataset, while AgentMove (Feng et al., 2025) leads on YJMob100K by leveraging its spatial memory module for the denser grid-based location space. LLM-Mob and TrajLLM show higher spatial error, with TrajLLM (Ju et al., 2025) particularly sensitive to the backbone's spatial reasoning capability (Dist. $>$ 6 km under GPT-4.1-mini). Despite each baseline's individual strengths, none simultaneously achieves high ranking accuracy and low spatial error across both datasets, while our method addresses through adaptive evidence gathering.

We further note that BW, which exhibits stronger temporal regularity (see Table 1), amplifies the advantage of end-to-end supervised models that can directly learn periodic patterns from training data, widening the gap between supervised and LLM-based approaches. Nevertheless, even on this more challenging setting, our framework still narrows the margin to within 6.74 percentage points of the best supervised baseline while maintaining full grounding on evidence.

To provide a more intuitive comparison, Figure 3 visualizes part of the predicted trajectories from a sample user. As shown in the figure, our method remains closer to the ground truth, whereas the baseline methods exhibit more noticeable spatial drift. This observation further supports the quantitative results in Table 2.

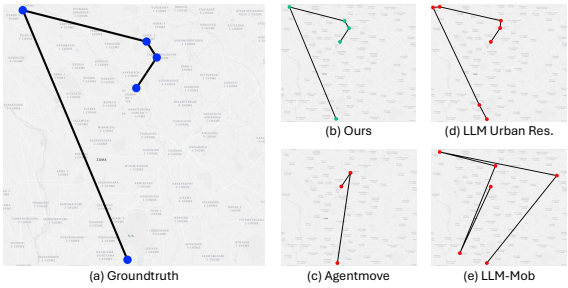

Figure 3: Part of predictions for a sample user. (a) Groundtruth (b) Our method (c-e) Baselines

4.2.2 Analysis of Reasoning Process

First, compared with prompt-engineering methods, our method treats next-location prediction more as an evidence-grounded decision problem than an open-ended prediction task. Instead of asking the large language model to complete the reasoning process inside the prompt, we turn this reasoning into structured evidence. We use several independent tools to model different types of information, such as Mobility Context Retriever , Geographical Information Retriever, and Stay-move Estimator. As a result, the prediction relies on a clear chain of evidence instead of mixed background knowledge inside the LLM. This design reduces the instability caused by prompt-based prediction. We provide Case 3 in Appendix B as case study. In this case, two nearby can-

didates receive conflicting temporal support, and the agent resolves the ambiguity through structured evidence: the stay-move estimator assigns 48% probability to the correct candidate versus 2% for the alternative, and the candidate ranker confirms this gap (0.450 vs. 0.080). To further verify, the agent proactively queries the location profiler and finds the competing candidate accounts for only 4.28% of historical visits — sporadic rather than habitual. In contrast, AgentMove predicts the wrong location based on proximity heuristics without cross-checking these statistics.

Second, we adopt an efficient and adaptive planning strategy, because not all samples require the same level of reasoning. Many trajectories have strong periodic patterns, and users frequently visiting the same locations at similar times of day. Our method predicts directly in a short path when the temporal pattern is clear and the initial evidence is consistent, otherwise calls agent tools to collect more clear evidence, which avoids unnecessary reasoning for simple cases.

When strong time regularities exist, our method achieves higher prediction accuracy than other agent-based approaches. We provide case 1 in Appendix B. In this case, the temporal pattern at the target hour strongly favors a single candidate (78% of historical same-day-and-hour instances), and the candidate ranking reflects this dominance with a score gap of 0.742 versus 0.082. The agent identifies the consistency across signals and skips further tool calls, producing the prediction in a single step. This fast path accounts for over 60% of all predictions in our experiments.

4.3 Ablations

4.3.1 Ablation Study on GPT-4.1-mini

Effect of the Stay-move Estimator As shown in Table 3, adding the Stay-move Estimator brings a small improvement in acc@1 on the tool path of our method. By studying on the full trace of agent inference, we found that it reduces both false stay and false move errors. Compared with the version without this tool, the false stay rate decreases by 2.49% in relative terms, and the false move rate decreases by 1.41% in relative terms. These results show that the Stay-move Estimator helps correct both types of boundary errors, with a larger effect on false stay. In other words, this tool is mainly helpful in correcting ambigous boundary cases. When the user should leave the current location, the tool helps push the model toward a move prediction. When the user should remain at the current location, the tool can also pull some predictions back to stay. We provide an example from the BW dataset below to illustrate how the Stay-move Estimator corrects a false-stay error.

Case:

user `007ee6b20600323fb29aa75d82015db3` at 2023-05-25 16:00

> **Without Stay-move Estimator**(*conf: 0.90*)
> *"Although the historical pattern shows the user at location 72837 at 16:00 on past Thursdays, the current context shows the user has been at location 72844 for the past 5 hours with low mobility and strong continuity. The top candidate with the highest score is also 72844. Therefore, the prediction favors location 72844 based on recent trajectory and stay state."*

> **With Stay-move Estimator**(*conf: 0.85*)
> *"Although the user is currently at location 72844 for 5 hours, historical patterns for the same day and hour (Thursday 16:00) show a dominant location at 72837. The transition probability from 72844 to 72837 at 16:00 is also high (37%), and 16:00 is a typical departure hour from 72844. Therefore, the prediction is a move to 72837 with high confidence based on historical day+hour patterns and transition probabilities."*

Effect of the Historical Behavior Retriever The Historical Behavior Retriever enables the agent to query location-specific visit statistics and transition patterns. Through analysis on full trace, removing the tool yields negligible overall accuracy changes (0.02% Acc@1 on BW, 0.13% on YJMob). It is selectively invoked on only 6.3% (BW) and 3.0% (YJMob) of predictions, targeting cases with substantially lower baseline accuracy (34.5% vs. 54.0% Acc@1 on BW), confirming that the agent appropriately identifies high-uncertainty scenarios for tool invocation. It helps and hurts in roughly equal measure (4.2% vs. 4.8% of invoked cases on BW).

The critical role of the tool is injecting confidence into the agent's reasoning grounded by

Table 3: Ablation study of different components on BW and YJMOB datasets. **Bolded** numbers indicate the best performance among all variants.

Backbone	Variant	BW			YJMOB		
		acc@1	MRR@5	Dist.	acc@1	MRR@5	Dist.
Qwen3-8b	Full	63.78%	74.91%	2.776	31.33%	45.57%	**4.359**
	w/o Stay-move Estimator	63.08%	74.56%	2.880	**31.40%**	45.62%	4.404
	w/o Historical Behavior Retriever	**64.02%**	**75.05%**	**2.765**	31.31%	**45.65%**	**4.359**
GPT-4.1-mini	Full	**67.44%**	**76.85%**	**2.412**	**32.37%**	**46.26%**	**4.219**
	w/o Stay-move Estimator	66.59%	76.41%	2.553	32.26%	46.22%	4.263
	w/o Historical Behavior Retriever	67.42%	76.85%	2.427	32.27%	46.21%	4.224

evidence. For example, without Historical Behavior Retriever, the agent predicts location 65117 but cites "no strong transition data" (conf=0.60); with it, the same prediction is supported by concrete evidence—33.5% visit frequency, 7.4-hour average stay, and 86.5% self-transition rate—raising confidence to 0.80. It thus shifts the agent from heuristic guessing to evidence-based reasoning, improving decision transparency without changing the outcome.

Case: user `e4a27b58` at 2023-05-26 13:00

> **Without HBR** (*conf: 0.60*)
> *"Although the user is currently at location 8761 with moderate mobility and no strong transition data, the top candidate by score is 65117, which aligns with the user's most visited location and day-hour patterns show a moderate probability for 65117. Historical Friday 13:00 data suggests 48795, but it ranks lower and has less visit frequency. The prediction favors the top candidate with high frequency and relevant scoring."*

> **With HBR** (*conf: 0.80*)
> *"Although the user is currently at location 8761, the top candidate by score and the user's most visited location is 65117. The historical weekly pattern shows a strong signal for location 48795 on Fridays at 13:00, but transition data shows no move to 48795 and current stay is just 1 hour at location 8761 with moderate mobility. Given the high visit frequency, score, and typical stay duration at 65117, it is likely the user will be at 65117 at the target time."*

4.3.2 Ablation Study on Qwen3-8b

Compared with GPT-4.1-mini, Qwen3-8b exhibits notably weaker multi-tool coordination. Agent trace analysis reveals that after entering the tool-calling mode, 29.1% of samples fail to invoke the Stay-Move Estimator as prescribed. We quantify the resulting loss as the product of the missed-call proportion and the accuracy drop on those samples, yielding an estimated loss of at least 0.08% Acc@1 on BW and 0.16% on YJMob100K. This unreliable invocation also explains a counter-intuitive result in Table 3: removing the Stay-Move Estimator actually yields the highest Acc@1 on YJMob100K for Qwen3-8b, because the tool's benefit is offset by the noise introduced when the model fails to call it consistently.

The root cause is Qwen3-8b's weaker instruction-following and tool-planning capability: it cannot reliably adhere to the prescribed tool-calling sequence, leading to broken evidence chains and incomplete integration of multiple evidence sources. In contrast, GPT-4.1-mini exhibits no obvious missed tool calls. These results indicate that the effectiveness of the agentic framework depends on the backbone model's tool-use competence—smaller models with limited instruction-following capacity cannot fully exploit the multi-tool reasoning pipeline, and may even be harmed by additional tools they cannot reliably operate.

5 Conclusion and Future Work

We proposed AgentMob, an efficient and evidence-grounded LLM-agent framework for mobility prediction that integrates adaptive tool-use with adaptive reasoning over historical mobility patterns. Future directions include (i) stronger uncertainty estimation and risk-aware planning, (ii) broader multi-city transfer, and (iii) richer human-in-the-loop controls.

6 Limitations

Our work has several limitations that suggest directions for future research. First, the framework's effectiveness is tied to the back-

bone LLM's tool-use competence. As the Qwen3-8b ablation shows, models with weaker instruction-following ability may struggle to fully exploit multi-tool reasoning, narrowing the performance gains. Adapting the framework for smaller or more efficient models—e.g., through distillation of tool-use behavior or simplified orchestration protocols—remains an open challenge.

Finally, the current framework employs a fixed toolbox with manually defined functional boundaries and a hand-crafted coordination protocol. While effective in the prediction scenarios studied here, both the functional decomposition of individual tools and the inter-tool collaboration pattern (invocation order, triggering criteria, fast-path thresholds) have room for systematic optimization. Jointly refining tool design and orchestration—e.g., via automated prompt tuning, reinforcement-learning-based tool selection, or dynamic tool composition—is a promising direction for future work.

References

Hugo Barbosa, Fernando B. de Lima-Neto, Alexandre Evsukoff, and Ronaldo Menezes. 2020. The scales of human mobility. *Nature*, 586(7831):402–407.

Blogwatcher, Inc. 2024. Blogwatcher. https://www.blogwatcher.co.jp/. (Japanese only).

Christian Calderón, Pasqual Martí, Jaume Jordán, Javier Palanca, and Vicente Julian. 2025. Cognitive agents in urban mobility: Integrating llm reasoning into multi-agent simulations. *Sensors*, 25(18):5688.

Jie Feng, Yuwei Du, and Yong Li. 2024. Limp: Large language model enhanced intent-aware mobility prediction. *arXiv preprint arXiv:2408.12832*.

Jie Feng, Yuwei Du, Jie Zhao, and Yong Li. 2025. AgentMove: A large language model-based agentic framework for zero-shot next location prediction.

Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 world wide web conference*, pages 1459–1468.

Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2012. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, MPM '12, New York, NY, USA. Association for Computing Machinery.

Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Qiguang Chen, and 1 others. 2025. Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation. *arXiv preprint arXiv:2505.23885*.

Chenlu Ju, Jiaxin Liu, Shobhit Sinha, Hao Xue, and Flora Salim. 2025. TrajLLM: A modular LLM-enhanced agent-based framework for realistic human trajectory simulation.

Guohao Li and 1 others. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *arXiv preprint arXiv:2303.17760*.

Peibo Li, Maarten de Rijke, Hao Xue, Shuang Ao, Yang Song, and Flora D. Salim. 2024a. Large language models for next point-of-interest recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 1463–1472. ACM.

Qiumeng Li, Chunhou Ji, and Xinyue Liu. 2025. From narrative to action: A hierarchical llm-agent framework for human mobility generation. *Preprint*, arXiv:2510.24802.

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

Defu Lian, Yongji Wu, Yong Ge, Xing Xie, and Enhong Chen. 2020. Geography-aware sequential location recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2009–2019.

Yuebing Liang, Yichao Liu, Xiaohan Wang, and Zhan Zhao. 2024. Exploring large language models for human mobility prediction under public events. *Computers, Environment and Urban Systems*, 112:102153.

Qi Liu, Can Li, and Wanjing Ma. 2025. Gatsim: Urban mobility simulation with generative agents. *arXiv preprint arXiv:2506.23306*.

Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2016. Predicting the next location: a recurrent model with spatial and temporal contexts. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 194–200. AAAI Press.

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

Yingtao Luo, Qiang Liu, and Zhaocheng Liu. 2021. Stan: Spatio-temporal attention network for next location recommendation. In *Proceedings of the web conference 2021*, pages 2177–2185.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Karl Cobbe, Liane Dao, Matthew Jones, Nicholas Kornis, Vlad Malaya, Kathryn Millican, Pamela Mishkin, and 1 others. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Nuria Oliver, Bruno Lepri, Harald Sterly, and 1 others. 2020. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Science Advances*, 6(23):eabc0764.

OpenAI. 2025. Gpt-4.1-mini: A compact and efficient large language model. https://openai.com/. Accessed: 2026-03-16.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Joon Sung Park, Carrie O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Canwen Qin, Aston Zhang, Zhuosheng Chen, Heng Ji, Xiang Ren, Yizhou Sun, and 1 others. 2023. Toolllm: Facilitating large language models to master 16,000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Yanjun Qin, Yuchen Fang, Haiyong Luo, Fang Zhao, and Chenxing Wang. 2022. Next point-of-interest recommendation with auto-correlation enhanced multi-modal transformer network. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2612–2616.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820.

Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. 'smolagents': a smol library to build great agentic systems. https://github.com/huggingface/smolagents.

S. Sengupta and 1 others. 2024. Mag-v: Multi-agent generative verification for robust data synthesis. *arXiv preprint arXiv:2412.04494*.

Tianao Sun, Ke Fu, Weiming Huang, Kai Zhao, Yongshun Gong, and Meng Chen. 2024. Going where, by whom, and at what time: Next location prediction considering user preference and temporal regularity. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2784–2793.

Y. Tang and 1 others. 2025. Agent-kb: Knowledge base enhanced multi-agent collaboration for complex task solving. *arXiv preprint arXiv:2507.06229*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

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

Xinglei Wang, Meng Fang, Zichao Zeng, and Tao Cheng. 2023. Where would i go next? large language models as human mobility predictors. *arXiv preprint arXiv:2308.15197*.

Xinglei Wang, Meng Fang, Zichao Zeng, and Tao Cheng. 2024b. Where would i go next? large language models as human mobility predictors. *arXiv preprint arXiv:2308.15197*.

Yu Wang, Tongya Zheng, Shunyu Liu, Zunlei Feng, Kaixuan Chen, Yunzhi Hao, and Mingli Song. 2024c. Spatiotemporal-augmented graph neural networks for human mobility simulation. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):7074–7086.

Jiaman Wu, Shangqing Cao, Giuseppe Perona, and Marta C Gonzalez. 2024. Imitate the right data: City-wide mobility generation with graph learning. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, pages 609–612.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, Yoshihide Sekimoto, Kaoru Sezaki, Esteban Moro, and Alex Pentland. 2024. Yj-mob100k: City-scale and longitudinal dataset of

anonymized human mobility trajectories. *Scientific Data*, 11(1):397.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and Others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *Preprint*, arXiv:2401.02385.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2026. A survey of large language models. *Preprint*, arXiv:2303.18223.

Yu Zheng. 2015. Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology*, 6(3):1–41.

Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology*, 5(3):1–55.

Lin Zhong, Lingzhi Wang, Xu Yang, and Qing Liao. 2025. Comapoi: A collaborative multi-agent framework for next poi prediction bridging the gap between trajectory and language. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1768–1778.

A Location Description prediction

Since raw GPS coordinates and anonymous cell IDs carry no semantic information, we design a description-prediction step that converts each spatial unit into a concise textual characterization, enabling the LLM to reason about locations based on their functional roles. All descriptions are generated once with GPT-4o and shared across every LLM-based method evaluated in this paper.

BW. Each polygon corresponds to a third-level administrative unit and is identified by a hierarchical place name (prefecture / municipality / district, e.g., *Tōkyō-to, Itabashi-ku, Hasune*). We retrieve publicly available geographic information for each place name and prompt GPT-4o to produce a one-sentence functional summary. For example, the polygon *Tōkyō-to, Itabashi-ku, Hasune* yields:

> *"It functions primarily as a quiet residential neighborhood, where the area is characterized by apartment complexes, small parks, and local convenience stores, with nearby transit access via Hasune Station on the Mita Line."*

YJMob100K. Cells are anonymized $500\,\mathrm{m} \times 500\,\mathrm{m}$ grid squares with no place names. We leverage the POI category vectors (85 categories) provided by the dataset authors (Yabe et al., 2024), and prompt GPT-4o to summarize each cell's top POI categories and their relative proportions into a one-sentence description. For example, a cell whose top categories are *Hair Salon* (8.3%), *Transit Station* (7.4%), and *Hospital* (7.0%) yields:

> *"It functions primarily as a diverse mixed-use service area, where the mapped profile is led by many hair salons alongside many transit stations and many hospitals, with adjacent cells indicating a more active, service-rich amenity landscape and no single activity clearly dominating the amenity mix."*

B Inference Case Studies

Our case studies reveal two typical inference patterns of the proposed method. When strong day-hour regularity exists, the model can directly rely on structured temporal evidence and make accurate predictions through the fast-path mechanism. When the temporal signal is not sufficient for direct fast-path prediction, the model performs multi-step reasoning by explicitly querying structured tools and resolving conflicts across different evidence sources. In contrast, AgentMove more often relies on broader semantic heuristics such as the current location, the most visited location, or generic commuting assumptions, which can lead to incorrect predictions.

B.1 Case 1: Fast-path under strong temporal regularity

Dataset:BW
Ground truth: 36166
Ours: correct
AgentMove: wrong (17844 ranked first)

Ours inference

> The prediction is made directly through the fast-path mechanism.

The temporal evidence shows a perfectly repeated pattern: on Friday at 09:00, the user was at location 36166 in 18 out of 18 historical instances. Since the day-hour pattern is fully consistent, the model can confidently rely on this structured temporal signal without invoking additional tools.

AgentMove inference

The user's most visited location is 17844 with a high overall visit rate and a significant presence in their long-term history, indicating a strong likelihood to return especially during weekday mornings.
... 36166 ... is ranked third.

B.2 Case 2: Multi-step reasoning with an additional verification step

Dataset:BW
Ground truth: 72883
Ours: correct
AgentMove: wrong (35990 ranked first)

Ours inference

The agent first called `Candidate Location Ranker`, `Mobility Context Retriever`, and `Stay Move Estimator` to collect complementary evidence from candidate ranking, temporal context, and current mobility state. The candidate ranking did not strongly favor the final answer: 72883 was not the most obvious top choice under the initial composite score, and other candidates such as 35990 appeared more competitive under general ranking signals. However, the prediction context showed that the current query occurred in a highly regular temporal slot, and the same day-hour historical pattern was much more consistent with 72883 than with the other candidates. At the same time, the stay/move evidence suggested that the user was at a departure-ready state rather than in a stable long stay, so the decision should not rely only on the current location.

Instead of submitting a prediction immediately, the agent performed an additional verification step by calling `Location Profiler` for 72883. This extra query returned more detailed structured evidence about the candidate, including its historical visitation characteristics and its compatibility with the current temporal context. Based on this full evidence chain, the agent explicitly concluded that although the candidate score ranking was lower, the historical temporal pattern was a strong override. Therefore, the model did not follow the ranking mechanically. It used the additional tool result to verify that 72883 was the most plausible destination under the current temporal condition, and finally predicted 72883.

AgentMove inference

AgentMove ranked 35990 first and relied mainly on broader semantic and frequency-based heuristics. Its reasoning focused on the current location context, high-frequency familiar areas, and the general assumption that users tend to remain within or return to commonly visited nearby regions. Although 72883 was compatible with the user's history, AgentMove did not explicitly verify the stronger day-hour regularity associated with that location, nor did it perform an additional structured check to resolve the conflict between temporal evidence and general ranking intuition. As a result, it stayed with the more generic high-frequency candidate 35990 and missed the correct answer.

B.3 Case 3: Multi-step reasoning under conflicting evidence

Dataset:YJMob100K
Ground truth: (132,46)
Ours: correct
AgentMove: wrong ((132,47) ranked first)

Ours inference

The agent first called `Stay Move Estimator`, `Candidate Location Ranker`, and `Mobility Context Retriever` to determine whether the user was likely to leave the current location, which destinations were most plausible, and whether there was any strong temporal regularity. The prediction context showed that the user was currently at (132,46) and had already stayed there for 4 hours. The recent trajectory was also stable, with the user remaining at (132,46) for the last several time steps. At the same time, the temporal statistics were conflicting: the hourly pattern at 07:00 favored (132,46) with 46.3%, while some matched weekday examples suggested (132,47) as a possible destination. The context tool explicitly warned that this temporal signal should be verified against transition-based evidence before overriding the current state.

The stay/move tool provided more fine-grained evidence. It showed that although the current stay duration of 4 hours was longer than the historical average of 2.0 hours, the hour-specific transition distribution from (132,46) at 07:00 still assigned the highest probability to staying at (132,46) with 48%, followed by (131,46) with 39%, while (132,47) accounted for only 2%. Consistently, the candidate ranking placed (132,46) first with score 0.450, (131,46) second with 0.251, and (132,47) only fourth with 0.080. Therefore, the evidence at this stage suggested that the temporal clue toward (132,47) was not sufficient to outweigh the stronger transition and continuity evidence supporting (132,46).

To further resolve this conflict, the agent called `Location Profiler` for (132,47). This tool showed that (132,47) had only 7 historical visits in total, accounting for 4.28% of all visits, and therefore did not represent a dominant destination in the user's routine. In addition, its own transition profile suggested that visits to (132,47) were relatively sporadic rather than part of a stable pattern. This additional evidence clarified that the apparent support for (132,47) came from a small number of matched temporal examples, rather than from a strong and consistent mobility pattern.

Combining the stable recent trajectory, the high stay probability at 07:00, the top-ranked candidate score of (132,46), and the low-support diagnostic for (132,47), the agent finally predicted (132,46).

AgentMove inference

The user's current location is (132,46), which is also one of the most frequently visited places in the long-term history. Based on recent movement patterns and general mobility regularity, (132,47) is considered a plausible next location because it appears in temporally similar contexts and is geographically close to the current position. The prediction is further supported by the assumption that users often move between nearby routine locations during morning hours. Therefore, (132,47) is ranked first, while (132,46) remains another strong candidate due to its overall visit frequency and current occupancy.