Why Users Go There: World Knowledge-Augmented Generative Next POI Recommendation

Qiuyu Ding[1], Heng-Da Xu[1], Wei Zhang[1], Dongyi Lv[1,2], Changda Xia[1], Feng Xiong[1], Mu Xu[1]*,

[1]Amap, Alibaba Group [2]Xi'an Jiaotong University,

{qiuyu.dqy,xuhengda.xhd,chengxin.zw}@alibaba-inc.com

{xiachangda.xcd,xf250971,xumu.xm}@alibaba-inc.com

lvdongyi@stu.xjtu.edu.cn

Abstract

Generative point-of-interest (POI) recommendation models based on large language models (LLMs) have shown promising results by formulating next POI prediction as a sequence generation task. However, the knowledge encoded in these models remains fixed after training, making them unable to perceive evolving real-world conditions that shape user mobility decisions, such as local events and cultural trends. To bridge this gap, we propose AWARE (Agent-based World knowledge Augmented REcommendation), which employs an LLM agent to generate location- and time-aware contextual narratives that capture regional cultural characteristics, seasonal trends, and ongoing events relevant to each user. Rather than introducing generic or noisy information, AWARE further anchors these narratives in each user's behavioral context, grounding external world knowledge in personalized spatial-temporal patterns. Extensive experiments on three real-world datasets demonstrate that AWARE consistently outperforms competitive baselines, achieving up to 12.4% relative improvement.

1 Introduction

With the proliferation of location-based social networks (LBSNs) and mobile devices, massive volumes of user check-in data have become available, enabling the study of human mobility at an unprecedented scale (Gonzalez et al., 2008). A central task in this domain is *next point-of-interest (POI) recommendation*: given a user's historical check-in trajectory, predicting the most likely next location they will visit (Li et al., 2024; Feng et al., 2024a; Halder et al., 2022; Zhong et al., 2025). Accurate next POI prediction directly benefits real-world applications, including personalized navigation, location-aware advertising, and urban planning (Zheng, 2015), and has attracted increasing attention from both academia and industry in recent years.

*Corresponding author.

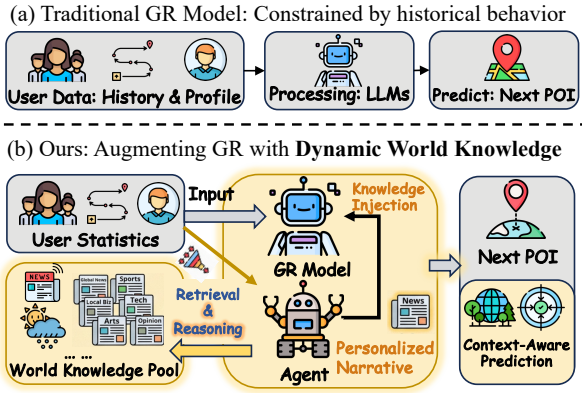

Figure 1: Comparison of (a) traditional generative recommendation using only historical check-in data, and (b) AWARE, which augments the GR model with dynamic world knowledge selected by an LLM agent.

Existing approaches have evolved from collaborative filtering and Markov chain models (Feng et al., 2015) to recurrent and attention-based architectures (Hochreiter and Schmidhuber, 1997; Luo et al., 2021; Yang et al., 2022; Yan et al., 2023) that capture increasingly complex spatial-temporal dependencies. More recently, large language models (LLMs) (Ding et al., 2026; Rajput et al., 2023; Chen et al., 2025; Jiang et al., 2026; Zhou et al., 2025) have given rise to *generative POI recommendation* (Li et al., 2024; Wang et al., 2025; Liu et al., 2025; Lv et al., 2026), which serializes check-in history into a natural language prompt and fine-tunes an LLM to directly generate the next POI. By exploiting rich textual representations rather than opaque ID embeddings, generative approaches achieve strong performance on next POI prediction benchmarks.

Despite their promise, existing generative POI recommendation models share a critical limitation: once fine-tuned on historical data, they operate as **static systems** whose knowledge is entirely frozen at training time. This renders them unable to capture evolving real-world context, such as local

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 50446–50459
October 24-29, 2026 ©2026 Association for Computational Linguistics

events, cultural trends, and seasonal shifts, that fundamentally shapes user mobility decisions. For example, a user who routinely visits a gym on weekday evenings may instead attend a community fitness event at a nearby park; a static model, unaware of this event, would confidently predict the gym and fail. Without access to such external context, existing methods capture *where* users have been but not *why* they go next.

To address this gap, we propose **AWARE** (**A**gent-based **W**orld-**A**ugmented **RE**commendation), a framework that injects dynamic, personalized world knowledge into the generative recommendation pipeline (Figure 1). For each user's activity region and time period, AWARE employs an LLM agent to synthesize contextual narratives, termed *hotspot text*, describing regional cultural characteristics, seasonal trends, and ongoing events. Unlike RAG, which returns pre-indexed documents irrespective of user identity, our agent reasons over each user's mobility history to produce personalized narratives that distill relevant world context into actionable conclusions. To ensure the injected knowledge remains user-specific rather than generic, AWARE further anchors the hotspot text in behavioral signals such as frequently visited locations and habitual spatial transitions.

2 Background

2.1 Problem Formulation

Let $\mathcal{U}$ denote the set of users and $\mathcal{P}$ the set of POIs. Each POI $p \in \mathcal{P}$ is associated with a name, a category, and geographic coordinates. A check-in (u, p, t) records that user u visited POI p at time t. For each user u, the check-in history is an ordered sequence $\mathcal{S}_u = [(p_1, t_1), (p_2, t_2), \ldots, (p_n, t_n)]$. The task is to predict the next POI p_{n+1} given $\mathcal{S}_u$ and the current time context.

2.2 Generative POI Recommendation

Recent work reframes next POI prediction as a text generation task (Li et al., 2024; Wang et al., 2025; Liu et al., 2025; Lv et al., 2026). A user's check-in history is serialized into a structured prompt containing POI names, categories, timestamps, and geographic information. An LLM is then fine-tuned to generate the next POI as its output sequence.

Formally, a prompt construction function $\phi(\cdot)$ converts $\mathcal{S}_u$ into a textual input $x = \phi(\mathcal{S}_u)$, and the model is trained to maximize $P(y \mid x)$, where y is the ground-truth next POI. This formulation enables the model to leverage pre-trained semantic knowledge about location names and categories, providing richer representations than ID-based.

2.3 Semantic ID

Rather than representing POIs by raw names, which consume lengthy token sequences and lack structural relationships, recent work assigns each POI a compact *Semantic ID* (SID) (Wang et al., 2025). SID encodes each POI as a short compositional token sequence whose components capture different facets and can be shared across POIs:

```
Coffee Shop, 1128 3rd Ave  ⇒  <a_29><b_31><c_20>
Coffee Shop, 5 W 21st St   ⇒  <a_26><b_3><c_20>
```

Both POIs share `<c_20>` due to their common category, while the remaining tokens capture individual identity. AWARE is agnostic to the POI representation scheme; in this work we adopt a SID-equipped generative backbone (details in Section 3).

2.4 World Knowledge for POI Recommendation

In this work, world knowledge refers to dynamic, location-specific contextual information beyond check-in records, including regional cultural characteristics, local events, and seasonal trends. Such knowledge evolves continuously and often drives user mobility decisions, yet cannot be encoded statically within model parameters.

We define world knowledge as $w = \mathcal{A}(\mathcal{S}_u, l, t)$, where $\mathcal{A}$ denotes an LLM agent, l is the user's activity region, and t is a temporal context. Unlike generic knowledge retrieval, $\mathcal{A}$ first analyzes the user's mobility patterns from $\mathcal{S}_u$, then incorporates real-world events and trends relevant to region l at time t. For example, given a user who commutes between Queens and Midtown Manhattan and has visited Yankee Stadium on weekends during April–May 2012, the agent may generate:

> *"The user exhibits a commuting pattern between Queens and Midtown Manhattan. Combined with the Yankees' home season in April–May, the user is likely interested in upcoming home games or nearby venues."*

We term this output *hotspot text*, a personalized narrative that bridges user behavior with external world dynamics to reason about why a user is likely to visit certain POIs.

3 Methodology

In this section, we present the AWARE framework (see Figure 2). We first describe the personalized

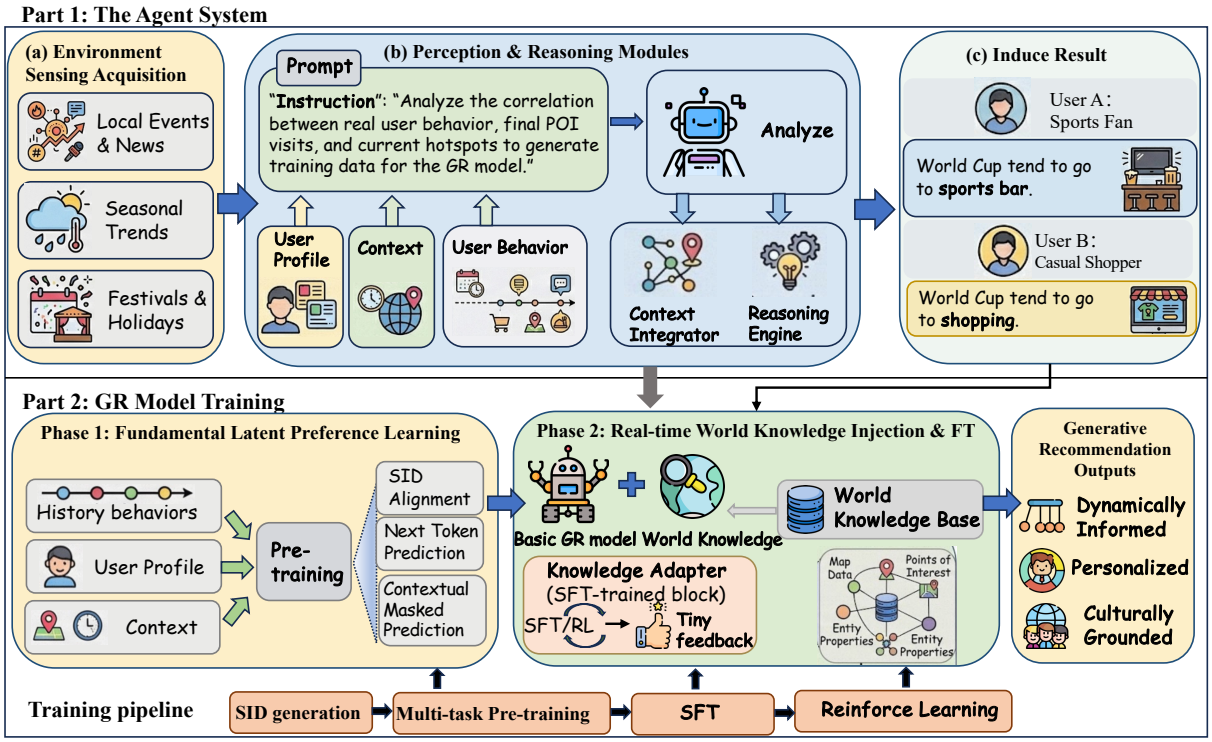

Figure 2: Overview of the AWARE framework. (1) An LLM agent retrieves external signals (e.g., news, events, festivals) and integrates them with user behavioral patterns to produce personalized narratives. (2) The pretrained GR model is fine-tuned with the acquired world knowledge to generate context-aware recommendations.

world knowledge acquisition (Section 3.1), then introduce user-aware knowledge alignment (Section 3.2), and finally detail the context-aware generative inference (Section 3.3).

3.1 World Knowledge Acquisition Agent

Existing generative POI models treat knowledge as static, yet user mobility is shaped by evolving real-world dynamics that no fixed-parameter model can capture. External sources such as news feeds, event calendars, and travel portals offer a natural remedy, but their content is vast and largely irrelevant to any individual user. This motivates AWARE's design principle: leveraging each user's behavioral patterns as an anchor for world knowledge selection, so that the distilled signals capture not only *where* users have been but *why* they are likely to go next.

3.1.1 Three-Stage Reasoning Strategy

To this end, AWARE instantiates a tool-augmented LLM agent $\mathcal{A}$ that, given a user's trajectory $\mathcal{S}_u$ and the temporal–spatial context (t, l) implied by the tail of the trajectory, is equipped with a retrieval toolbox $\mathcal{R} = \{\texttt{web_search}, \texttt{web_fetch}\}$ and produces a compact hotspot text:

$$w = \mathcal{A}\big(\mathcal{S}_u,\, t,\, l\,;\, \mathcal{R},\, \Pi\big), \qquad (1)$$

where Π denotes the system prompt that governs the agent's reasoning protocol. The protocol decomposes into three sequential stages, with a detailed description provided below.

Stage (i): Behavioral Profile Inference. The agent first parses $\mathcal{S}_u$ in chronological order and weights the tail of the sequence more heavily to identify the user's active neighborhood l, recurring interests (e.g., dining style, commute routine, fitness habits, nightlife preferences), and plausible near-term intent. This profile is retained as an internal anchor and never exposed verbatim in the final output, ensuring that downstream retrieval is driven by user-specific priors rather than generic demographic assumptions.

Stage (ii): Constrained External Retrieval. Conditioned on the inferred profile, the agent formulates targeted web queries that jointly bind three constraints: *geographic* (restricted to l and its surrounding neighborhoods), *temporal* (anchored at t, with the target year and month embedded explicitly in every query), and *thematic* (the inferred interest categories). Retrieval proceeds iteratively: broader `web_search` calls first surface candidate event calendars, neighborhood guides, and news items, after

which `web_fetch` verifies specific claims on the most informative pages. The agent is explicitly instructed to corroborate each substantive claim across at least two independent sources and to discard any signal that cannot be anchored to the target time window, which mitigates hallucinations inherited from the LLM's parametric memory.

Stage (iii): Grounded Synthesis. In the final stage, the agent weaves the internal behavioral anchor together with the verified external signals into a single coherent paragraph w. The paragraph is required to interleave three elements in a tightly integrated manner: (a) what the user's history suggests, (b) what regional context is salient at time t, and (c) what POI categories or venue styles the user is most likely to seek next. Outputs that merely recite the trajectory, or that read as generic city guides disconnected from user behavior, are rejected by the format constraints in Π, ensuring that w reads as a *user-grounded contextual interpretation* rather than a document summary.

3.1.2 Agent Instantiation and Prompt Design

We instantiate $\mathcal{A}$ with a large-capacity LLM (`Qwen3-max`) augmented with a web-retrieval toolbox $\mathcal{R}$ that provides date-filtered web search and page-level content extraction. The system prompt Π encodes the three-stage protocol as explicit instructions and imposes three key constraints: (i) all factual claims must be grounded in retrieved evidence rather than parametric memory, (ii) the output must be a single coherent paragraph, and (iii) the length must not exceed a word budget M, which can be adjusted according to application requirements. Figure 3 visualizes the key components of Π, including the quality safeguards elaborated in Section 3.1.3. The complete system prompt is provided in Appendix C.

3.1.3 Generation Quality Control

Beyond the reasoning protocol, AWARE enforces three operational constraints on agent outputs, namely *output-format control*, *tool-usage efficiency*, and *temporal fidelity*, each implemented across prompt, tool, and post-hoc validation layers to forestall any single failure mode.

Output-Format Control. Since the downstream GR model is prompt-length sensitive, we enforce the word budget M via a prompt, validator, and truncator cascade: Π instructs the agent to cap w at M words and self-check length; if the validator detects $|w| > M$, the agent is re-invoked with a brief rewrite instruction; any residual overshoot is hard-truncated. We instantiate two budgets, $M = 80$ and $M = 150$, which anchor the informativeness versus noise trade-off in Section 4.5.

System Prompt Π of the Knowledge Agent

Role. The agent for next POI recommendation in city l.

Input. User trajectory $\mathcal{S}_u$: {(time, POI, type, address, lat, lon)}.

Tools. `web_search` (with publication-date filter) and `web_fetch`.

Protocol.

(1) Infer user preferences, routines, and near-term intent from $\mathcal{S}_u$ (tail-weighted).

(2) Identify the active region and activity context.

(3) Issue parallel queries for region- and time-local hotspots, events, and trends.

(4) Synthesize one paragraph weaving behavior, context, and predicted POI categories.

Temporal fidelity. Treat the latest timestamp in $\mathcal{S}_u$ as "now"; embed the target year/month in every query; prefer archival sources; drop unverifiable claims.

Search budget. At most two retrieval rounds; emit parallel queries rather than serial refinements.

Output. One compact English paragraph ($\leq M$ words); no structured formatting or trajectory recitation.

Figure 3: Condensed view of the system prompt Π. Beyond the reasoning protocol, Π encodes runtime guarantees for temporal fidelity, search-budget control, and output-length control.

Tool Usage Efficiency. Since agent rollouts dominate preprocessing cost, Π caps retrieval at two rounds, a broad main round and an optional supplemental round for residual gaps, and encourages parallel rather than sequential queries within each round; our orchestrator dispatches them concurrently. A code-level guard hard-stops any attempt at a third round and forces synthesis, bounding per-user latency and cost.

Temporal Fidelity. As the evaluation benchmarks span 2009 to 2013 and predate contemporary LLM training corpora, AWARE grounds the agent in the user's own timeline: the latest timestamp in $\mathcal{S}_u$ is treated as "now," every query is anchored to the target year and month, and documents lacking period-appropriate corroboration are discarded, ensuring that the injected knowledge reflects the user's original temporal and spatial context.

3.2 User-Aware Knowledge Alignment

Indiscriminate injection of world knowledge risks introducing noise. To ensure that external context is actionable for each user, AWARE aligns world

knowledge with user-specific behavioral evidence through three complementary priors.

Frequency-Based Behavioral Prior. We compute a frequency-ranked list of each user's visited POIs:

$$\mathcal{F}_u = [(p_1, c_1, f_1), (p_2, c_2, f_2), \ldots] \quad (2)$$

where p_i denotes a POI, c_i its category, and f_i the visit count. This prior encodes long-term behavioral regularities and serves as a relevance filter: signals aligning with established patterns receive implicit reinforcement, while those lacking behavioral corroboration are down-weighted.

Transition-Based Spatial Grounding. We model user mobility dynamics by extracting POI-to-POI transition statistics. For each POI p_i visited by user u, we record the set of subsequent destinations and their frequencies:

$$\mathcal{T}_u(p_i) = \{(p_j, n_{ij})\}_{j=1}^{|\mathcal{N}(p_i)|} \quad (3)$$

where p_j is a destination POI and n_{ij} the transition count. Given the most recently visited POI, we select the top-k most frequent transitions as spatial grounding signals, channeling world knowledge through the user's movement structure.

Periodic Temporal Modeling. User mobility often exhibits strong *periodicity*: behavior on Monday mornings is more predictive of next Monday than activity from the preceding weekend. To capture such long-range periodic evidence, AWARE reserves a fraction β of the history budget for entries sharing the same day-of-week as the prediction target, filling the remaining slots with the most recent entries. Formally, given budget L and target day-of-week d, we select $\lfloor \beta L \rfloor$ entries from $\{e \in \mathcal{S}_u \mid \text{dow}(e) = d\}$ and $(L - \lfloor \beta L \rfloor)$ recent entries from the remainder. This ensures that temporal knowledge from the agent (e.g., "Friday evening trends") is paired with the user's corresponding historical behavior.

3.3 Context-Aware Generative Inference

Before describing how AWARE constructs its input, we briefly outline the generative recommendation (GR) pipeline. Each POI is assigned an SID by embedding its textual description through a pretrained language model and discretizing via hierarchical clustering. A geospatial prefix derived from S2 cells (Google Inc., 2023) is prepended to encode geographic locality at multiple resolutions, enabling spatially proximate POIs to share prefix tokens. The model is trained to generate the ground-truth next POI's SID given the user's historical sequence, and autoregressively produces the predicted SID at inference time (Lv et al., 2026).

Building on this pipeline, AWARE organizes the acquired knowledge (Sec. 3.2) into a structured prompt providing three complementary dimensions beyond the check-in sequence:

$$x = \phi(\mathcal{H}_u, \mathcal{F}_u, w, \mathcal{C}_u, \mathcal{T}_u) \quad (4)$$

where $\mathcal{H}_u$ denotes the time-aware historical trajectories, $\mathcal{F}_u$ the POI visit frequency statistics, w the generated world knowledge, $\mathcal{C}_u$ the current trajectory to be continued, and $\mathcal{T}_u$ the transition pattern conditioned on the last visited POI. A complete prompt example is provided in Appendix B.

To accommodate the knowledge-enriched input, we fine-tune the backbone with LoRA (Hu et al., 2022), optimizing the generation objective over ground-truth next POI y given prompt x:

$$\mathcal{L} = -\log P_\theta(y \mid x) \quad (5)$$

Since AWARE operates entirely at the input representation level, it requires no architectural modification to the backbone and is orthogonal to downstream reasoning stages such as Chain-of-Thought and reinforcement learning, whose benefits can be combined additively.

4 Experiments and Analysis

4.1 Experimental Setting

Datasets. We evaluate AWARE on three real-world LBSN datasets that are widely adopted in next POI recommendation research: **Foursquare-NYC**, **Foursquare-TKY** (Yang et al., 2015), and **Gowalla-CA** (Cho et al., 2011). These datasets span diverse urban environments with varying scales of user activity and POI catalogs, providing a comprehensive testbed for evaluating recommendation methods across different mobility contexts. Detailed dataset preprocessing and statistics are summarized in Appendix A.

Baselines. To comprehensively assess the effectiveness of AWARE, we compare against representative baselines spanning three categories: (i) *Traditional* method: **PRME** (Feng et al., 2015), which learns metric embeddings for next-step transition prediction; (ii) *Neural-based* methods: **GETNext** (Yang et al., 2022), **TPG** (Luo et al., 2023), **MTNet** (Hu et al., 2018), **STHGCN** (Yan et al.,

Category	Method	NYC	TKY	CA
Traditional	PRME	0.1159	0.1052	0.0521
Neural-based	GETNext	0.2435	0.1829	0.1357
	TPG	0.2555	0.1420	0.1749
	MTNet	0.2620	0.2575	0.1453
	STHGCN	0.2734	0.2950	0.1730
	ROTAN	0.3106	0.2458	0.2199
LLM-based	SpaceTime-GR	0.2920	0.2610	0.1659
	LLM4POI	0.3372	0.3035	0.2065
	GNPR-SID	0.3618	0.3062	0.2403
	GA-LLM	0.3919	0.3482	0.2566
	CoAST	0.4027	0.3310	0.2721
	ROS	0.3925	0.3380	0.2703
Ours	**AWARE**	**0.4160**	**0.3599**	**0.3038**
	vs. ROS (Δ%)	+6.0%↑	+6.5%↑	+12.4%↑

Table 1: Next POI recommendation performance (HR@1). Best results are in **bold**. Δ% denotes the relative improvement of AWARE over baseline ROS.

Variant	HR@1	Δ%
AWARE (Full)	**0.4160**	–
World Knowledge Acquisition Agent		
- Grounded Synthesis (Stage 3)	0.4023	-3.3%
- Retrieval & Synthesis (Stage 2,3)	0.4040	-2.9%
- Whole World Knowledge (Stage 1,2,3)	0.4029	-3.1%
- Personality History Check-ins	0.4059	-2.4%
User-Aware Knowledge Alignment		
- Frequency-Based Behavioral Prior	0.4078	-2.0%
- Transition-Based Spatial Grounding	0.4071	-2.1%
- Periodic Temporal Modeling	0.4070	-2.2%
- All User Behavioral Evidences	0.4064	-2.3%
Without All Components (ROS)	0.3925	-5.6%

Table 2: Component ablation. Each row removes one component from the full model (HR@1 on NYC).

2023), and **ROTAN** (Feng et al., 2024b), which employ graph networks, attention mechanisms, or specialized spatio-temporal encoders to model sequential mobility patterns; and (iii) *LLM-based* methods: **SpaceTime-GR** (Lin et al., 2025), **LLM4POI** (Li et al., 2024), **GNPR-SID** (Wang et al., 2025), **GA-LLM** (Liu et al., 2025), **CoAST** (Zhai et al., 2025), and **ROS** (Lv et al., 2026), which reformulate next POI prediction as a sequence generation task using large language models with varying strategies for spatial reasoning. Detailed descriptions of all baselines are provided in Appendix D.

Evaluation Metrics. We adopt HR@1 as the primary evaluation metric, which measures whether the top-ranked prediction exactly matches the ground-truth POI. Besides, HR@K and NDCG@K for $K \in \{5, 10, 20\}$ are employed to further assess the ranking quality of the full candidate list. For each experiment, we run inference three times and report the mean score.

Implementation Details. AWARE adopts `Qwen3-4B` as the generative backbone with LoRA (Hu et al., 2022) fine-tuning ($r = 64$, $\alpha = 128$), and Spatial SIDs derived from `Qwen3-0.6B` embeddings. World knowledge is generated offline by an LLM agent (`Qwen3-max`) prior to training. Since the evaluation benchmarks span 2009–2013 and predate contemporary LLM training corpora, temporal fidelity is enforced via date-range filtering and a post-hoc temporal tolerance check ($\Delta = 30$ days) to prevent contamination. For time-aware history selection, we set the periodic ratio $\beta = 0.4$ and budget $L = 150$. Training runs for 2 epochs with learning rate 1e-4 on $8 \times$ AMD-Instinct-MI308X-OAM GPUs. We inherit all other hyperparameters from ROS (Lv et al., 2026). Full details of the agent setup and temporal fidelity enforcement are provided in Appendix E.

4.2 Main Results

Table 1 reports the next POI recommendation performance across three LBSN benchmarks. We adopt the ROS pretrained backbone as the controlled baseline, denoted as ROS throughout this paper, to isolate the contribution of world knowledge injection.

AWARE consistently outperforms ROS across all three datasets, achieving relative improvements of 6.0% on NYC, 6.5% on TKY, and 12.4% on CA. The particularly large gain on CA suggests that world knowledge is especially beneficial when check-in data is sparser, as external context can compensate for limited behavioral observations. Moreover, AWARE on NYC already exceeds strong LLM-based baselines including GA-LLM (0.3919) and CoAST (0.4027), despite these methods employing larger backbone models (7B parameters), indicating that enriching the input representation with grounded world knowledge can be more effective than scaling model capacity.

4.3 Ablation Study

Table 2 reports the ablation results on the NYC dataset, organized into two groups: the world knowledge acquisition agent and the user-level behavioral signals.

World Knowledge Acquisition Agent. The upper block of Table 2 investigates whether the gains stem from agent-based synthesis or merely from

Method	Easy (32%)	Hard (68%)	Overall
ROS	0.6911	0.2520	0.3925
AWARE	**0.6847**	**0.2896**	**0.4160**
Δ%	*-0.93%*	*+14.9%*	*+6.0%*

Table 3: Easy vs. hard case analysis (HR@1, NYC). Easy: ground truth appears in the user's last 5 visits.

richer prompt context. Removing all world knowledge yields a -3.1% drop, confirming its indispensable role. Notably, replacing synthesized text with raw search snippets causes an even larger drop (-3.3%), indicating that unprocessed retrieval introduces harmful noise. Retaining only parametric summarization (-2.9%) and generating non-personalized city-level descriptions (-2.4%) further confirm that external retrieval, agent-mediated synthesis, and user-level personalization are all necessary, and that the improvement cannot be attributed to prompt expansion alone.

User-Level Behavioral Signals. All three behavioral signals produce comparable degradation when removed (-2.0% to -2.2%), suggesting that each captures a distinct aspect of user mobility. The behavioral priors collectively account for substantial gains (-2.3%), indicating that grounding world knowledge in user-specific spatial-temporal patterns is essential for effective knowledge injection, as illustrated in our case study (Appendix F).

4.4 Generalization Beyond Familiar Patterns

A natural question is *where* the improvements concentrate. We partition the NYC test set into *easy* cases, where the ground-truth POI appears in the user's last 5 visits (32%), and *hard* cases where it does not (68%). As shown in Table 3, AWARE yields a substantial +14.9% relative gain on hard cases while easy cases remain largely stable (-0.93%). This asymmetry confirms that world knowledge and behavioral priors are most valuable when sequential patterns alone are insufficient: hard cases require reasoning beyond local trajectories, and the external context and long-range spatial regularities introduced by AWARE enable the model to generalize beyond familiar patterns and resolve inherently ambiguous predictions.

4.5 Hotspot Text Length Analysis

Since the granularity of world knowledge is controlled by the length of the hotspot text, we train separate models with the agent constrained to at most 80 and 150 words. As shown in Figure 4, world knowledge exhibits a clear *length–quality* relationship: the 80-word configuration (0.4050) offers only marginal improvement over the variant without hotspot text (0.4029 in Table 2), while the 150-word variant (0.4160) yields substantial and consistent gains across all three cities, with the improvement widening on sparser data (CA: +7.4% from 80 to 150 words). This validates the agent-based design: producing sufficiently grounded narratives requires contextual reasoning over user behavior and external signals, a capability that arises naturally from the LLM agent's generative process.

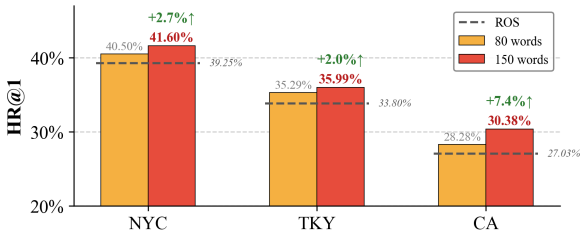

Figure 4: Effect of hotspot text length on HR@1. Maximum generation length is set to 80 and 150 words.

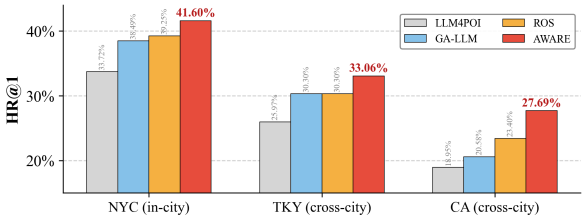

Figure 5: Cross-city generalization performance (trained on NYC). All results are HR@1.

4.6 Cross-City Generalization

To evaluate whether world knowledge facilitates transfer across cities, we train all models on NYC and directly evaluate on TKY and CA without any adaptation. As shown in Figure 5, AWARE consistently outperforms all baselines on both in-city and cross-city. We attribute this to the nature of world knowledge: the LLM agent encodes contextual information as natural language narratives describing general human mobility patterns, cultural habits, and temporal regularities, which are inherently not city-specific. This language-level generality allows AWARE to transfer meaningful contextual signals across geographic boundaries, complementing the structural transferability of the shared semantic space.

Metric	Traditional					Neural-based		LLM-based				
	SASRec	BERT4Rec	GRU4Rec	Caser	S^3-Rec	TPG	ROTAN	TIGER	GNPR-SID	ROS	ROS_{full}	AWARE
HR@5	0.3151	0.2857	0.1977	0.2883	0.3071	0.3551	0.4448	0.4965	0.5311	0.5791	0.6107	**0.6538**
HR@10	0.3896	0.3564	0.2460	0.3570	0.3854	0.4441	0.5223	0.5514	0.5942	0.6192	0.6563	**0.7173**
HR@20	0.4506	0.4130	0.2889	0.4135	0.4503	0.5121	0.5834	0.6001	0.6455	0.6420	0.6977	**0.7436**
NDCG@5	0.2224	0.2074	0.1442	0.2044	0.2235	0.2464	0.3471	0.4131	0.4430	0.5032	0.5355	**0.5483**
NDCG@10	0.2467	0.2304	0.1599	0.2267	0.2489	0.2755	0.3723	0.4276	0.4634	0.5163	0.5504	**0.5689**
NDCG@20	0.2622	0.2448	0.1708	0.2410	0.2654	0.2927	0.3878	0.4443	0.4766	0.5222	0.5608	**0.5759**

Table 4: Top-K evaluation on NYC. HR@K and NDCG@K with $K \in \{5, 10, 20\}$. ROS_{full} additionally incorporates Mobility CoT and spatial-guided RL. Best results are in **bold** and second best are underlined.

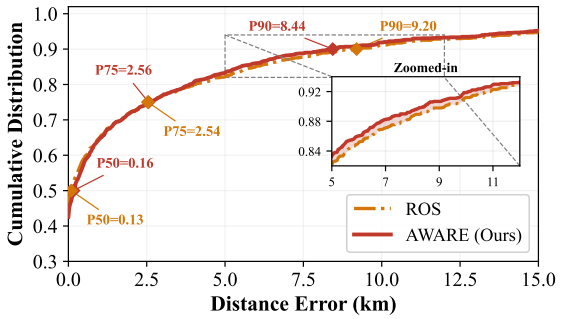

Figure 6: Cumulative distribution of prediction distance error on NYC. Larger area under the curve indicates better.

4.7 Top-K Recommendation Quality

Beyond top-1 accuracy, Table 4 reports HR@K and NDCG@K for $K \in \{5, 10, 20\}$ on NYC. AWARE surpasses not only ROS but also ROS_{full}, which additionally incorporates Mobility CoT and spatial-guided RL. The margin widens at higher K (e.g., HR@20: +15.8% over ROS), suggesting that world knowledge enriches the candidate space with contextually plausible POIs that reasoning-level enhancement alone cannot recover.

4.8 Prediction Distance Distribution

A natural concern is whether external knowledge introduces spatial noise. As shown in Figure 6, both methods perform comparably at P50 and P75, confirming that world knowledge does not degrade spatial precision. The difference emerges in the tail (P90: 8.44 vs. 9.20 km; mean: -3.4%), where AWARE produces fewer large-error predictions by resolving spatially ambiguous cases.

5 Related Work

Next POI Recommendation. Next POI recommendation has evolved from metric embeddings (Feng et al., 2015) and recurrent architectures (Hochreiter and Schmidhuber, 1997) to attention-based (Luo et al., 2021), graph-enhanced (Yang et al., 2022), and hypergraph methods (Yan et al., 2023). More recently, LLM-based approaches reformulate the task as sequence generation, with representative work including LLM4POI (Li et al., 2024), GNPR-SID (Wang et al., 2025), CoAST (Zhai et al., 2025), GA-LLM (Liu et al., 2025), and ROS (Lv et al., 2026).

Knowledge-Enhanced Recommendation. Prior POI recommendation work has leveraged geographic knowledge such as street-level addresses (Lv et al., 2026), category taxonomies (Wang et al., 2025), and spatial distance constraints (Liu et al., 2025). However, these sources are predominantly *static* and do not capture the dynamic real-world context that often drives user mobility decisions.

LLM Agents for Information Acquisition. Agent-based approaches such as RecMind (Wang et al., 2024) employ the LLM as the recommender itself, performing multi-step reasoning at inference time (Yao et al., 2023). While effective, this design couples reasoning cost with serving latency and forgoes the ability to learn fine-grained behavioral patterns from large-scale training data.

Different from methods above, AWARE uses the LLM agent solely for offline knowledge acquisition, allowing the generative backbone to internalize dynamic world knowledge through supervised learning and adapt it to individual user contexts.

6 Conclusion

We present AWARE, a framework that acquires world knowledge via an LLM agent, personalizes it through behavioral priors, and injects it into generative POI recommendation to capture context-dependent mobility intent. Experiments across multiple datasets show consistent gains over competitive baselines, and ablations confirm that the two signals are complementary, revealing not only where users have been but why they will go next.

Limitations

AWARE currently acquires world knowledge in an offline, pre-training stage: hotspot text is generated once per user and remains fixed throughout training and inference. While this design ensures zero additional latency at serving time, it cannot capture events that emerge after the knowledge acquisition pass. Extending the framework to support incremental or online knowledge refresh, where the agent periodically updates narratives in response to newly occurring events, would further improve temporal coverage. Additionally, the current agent relies on web search as its primary evidence source, which may offer uneven coverage across geographic regions and languages. Incorporating structured knowledge bases such as event calendars and transportation schedules could improve robustness in regions where web coverage is sparse.

References

Ben Chen, Xian Guo, Siyuan Wang, Zihan Liang, Yue Lv, Yufei Ma, Xinlong Xiao, Bowen Xue, Xuxin Zhang, Ying Yang, and 1 others. 2025. Onesearch: A preliminary exploration of the unified end-to-end generative framework for e-commerce search. *arXiv preprint arXiv:2509.03236*.

Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 1082–1090, New York, NY, USA. Association for Computing Machinery.

Yijie Ding, Jiacheng Li, Julian McAuley, and Yupeng Hou. 2026. Inductive generative recommendation via retrieval-based speculation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 14675–14683.

Shanshan Feng, Xutao Li, Yifeng Zeng, Gao Cong, Yeow Meng Chee, and Quan Yuan. 2015. Personalized ranking metric embedding for next new poi recommendation. In *IJCAI*, volume 15, pages 2069–2075.

Shanshan Feng, Haoming Lyu, Fan Li, Zhu Sun, and Caishun Chen. 2024a. Where to move next: Zero-shot generalization of llms for next poi recommendation. In *2024 ieee conference on artificial intelligence (cai)*, pages 1530–1535. IEEE.

Shanshan Feng, Feiyu Meng, Lisi Chen, Shuo Shang, and Yew Soon Ong. 2024b. Rotan: A rotation-based temporal attention network for time-specific next poi recommendation. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 759–770.

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

Google Inc. 2023. S2 geometry library. https://s2geometry.io/. Accessed: 2025-03-01.

Sajal Halder, Kwan Hui Lim, Jeffrey Chan, and Xiuzhen Zhang. 2022. Efficient itinerary recommendation via personalized poi selection and pruning. *Knowledge and Information Systems*, 64(4):963–993.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Guangneng Hu, Yu Zhang, and Qiang Yang. 2018. Mtnet: a neural approach for cross-domain recommendation with unstructured text. *KDD deep learning day*, pages 1–10.

Hao Jiang, Guoquan Wang, Donglin Zhou, Sheng Yu, Yang Zeng, Wencong Zeng, Kun Gai, and Guorui Zhou. 2026. Llm-aligned geographic item tokenization for local-life recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 14928–14936.

Dapeng Li. 2018. 1.08 - geocoding and reverse geocoding. In Bo Huang, editor, *Comprehensive Geographic Information Systems*, pages 95–109. Elsevier, Oxford.

Peibo Li, Maarten de Rijke, Hao Xue, Shuang Ao, Yang Song, and Flora D Salim. 2024. 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*, pages 1463–1472.

Haitao Lin, Zhen Yang, Jiawei Xue, Ziji Zhang, Luzhu Wang, Yikun Gu, Yao Xu, and Xin Li. 2025. Spacetime-gr: A spacetime-aware generative model for large scale online poi recommendation. *Preprint*, arXiv:2508.16126.

Zhao Liu, Wei Liu, Huajie Zhu, Jianxing Yu, Jian Yin, Wang-Chien Lee, and Shun Wang. 2025. Geography-aware large language models for next poi recommendation. *arXiv preprint arXiv:2505.13526*.

Yan Luo, Haoyi Duan, Ye Liu, and Fu-Lai Chung. 2023. Timestamps as prompts for geography-aware location recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1697–1706.

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.

Dongyi Lv, Qiuyu Ding, Heng-Da Xu, Zhaoxu Sun, Zhi Wang, Feng Xiong, and Mu Xu. 2026. Reasoning over space: Enabling geographic reasoning for llm-based generative next poi recommendation. *Preprint*, arXiv:2601.04562.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, and 1 others. 2023. Recommender systems with generative retrieval. *Advances in Neural Information Processing Systems*, 36:10299–10315.

Dongsheng Wang, Yuxi Huang, Shen Gao, Yifan Wang, Chengrui Huang, and Shuo Shang. 2025. Generative next poi recommendation with semantic id. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 2904–2914.

Yancheng Wang, Ziyan Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Yanbin Lu, Xiaojiang Huang, and Yingzhen Yang. 2024. Recmind: Large language model powered agent for recommendation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4351–4364.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Penglong Zhai, Jie Li, Fanyi Di, Yue Liu, Yifang Yuan, Jie Huang, Peng Wu, Sicong Wang, Mingyang Yin, Tingting Hu, and 1 others. 2025. Cognitive-aligned spatio-temporal large language models for next point-of-interest prediction. *arXiv preprint arXiv:2510.14702*.

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

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.

Guorui Zhou, Hengrui Hu, Hongtao Cheng, Huanjie Wang, Jiaxin Deng, Jinghao Zhang, Kuo Cai, Lejian Ren, Lu Ren, Liao Yu, and 1 others. 2025. Onerec-v2 technical report. *arXiv preprint arXiv:2508.20900*.

A Dataset Statistics

Following the established preprocessing pipeline of LLM4POI (Li et al., 2024), we: (i) filter users and POIs with fewer than 10 check-ins to ensure sufficient behavioral evidence; (ii) sort check-ins chronologically and segment them into trajectories using a 24-hour time-gap threshold; and (iii) split the data into 80%/10%/10% train/validation/test sets in temporal order, ensuring that all validation and test users and POIs have been observed during training. We additionally apply reverse geocoding (Li, 2018) to obtain street-level addresses as auxiliary textual cues.

Dataset	Users	POIs	Trajs	Categories	Check-ins
NYC	1,048	4,981	14,130	318	103,941
TKY	2,282	7,833	65,499	291	405,000
CA	3,957	9,690	45,123	296	238,369

Table 5: Dataset statistics after preprocessing. Two consecutive check-ins are assigned to the same trajectory if their time gap is within 24 hours.

B Prompt Example

Table 6 presents a complete input–output example from the NYC dataset. The input prompt consists of four components:

(i) `<user_poi_stats>` encoding the frequency-based behavioral prior

(ii) `<transition_patterns>` encoding spatial transition patterns

(iii) <user_preference> containing the world knowledge narrative generated by the LLM agent

(iv) the chronological check-in sequence with SID representations.

The model generates the SID of the predicted next POI as output.

C World Knowledge Acquisition Agent Prompt

Table 7 presents the complete system prompt used to instantiate the world knowledge acquisition agent. The prompt defines the agent's role, reasoning protocol, tool usage requirements, and output constraints. City name and maximum word count are adapted per dataset.

D Baseline Descriptions

We provide detailed descriptions of all baseline methods compared in our experiments.

PRME. PRME (Feng et al., 2015) learns user and POI embeddings in a shared metric space, where next-step transitions are predicted by ranking candidate POIs based on their distance to the user's current state. It captures sequential transition patterns through pairwise learning-to-rank objectives.

GETNext. GETNext (Yang et al., 2022) employs a graph-enhanced transformer architecture for next POI recommendation. It constructs a global POI transition graph to capture collaborative transition patterns and integrates spatial and temporal information through dedicated embedding layers, combining graph-level structural signals with transformer-based sequential modeling.

TPG. TPG (Luo et al., 2023) incorporates fine-grained temporal information into POI recommendation by encoding timestamps as continuous features. It models the temporal regularity of user visits and combines temporal signals with spatial and categorical features for next POI prediction.

MTNet. MTNet (Hu et al., 2018) is a multi-task learning framework that jointly models next POI prediction, category prediction, and arrival time estimation. By sharing representations across related tasks, it captures richer user mobility patterns than single-task approaches.

STHGCN. STHGCN (Yan et al., 2023) models higher-order spatial-temporal correlations among users and POIs through hypergraph convolutions. It constructs hyperedges that connect multiple POIs visited within the same trajectory, capturing group-level mobility patterns beyond pairwise transitions.

ROTAN. ROTAN (Feng et al., 2024b) employs a rotation-based attention mechanism to model spatial relationships in next POI recommendation. It encodes geographic coordinates as rotation angles in the attention computation, enabling the model to directly reason about spatial proximity and directional patterns in user trajectories.

SpaceTime-GR. SpaceTime-GR (Lin et al., 2025) is a space-time aware generative recommendation model that incorporates spatial and temporal encodings into the generative framework. It discretizes geographic coordinates and timestamps into token sequences, allowing the language model to jointly generate spatial-temporal-aware POI predictions.

LLM4POI. LLM4POI (Li et al., 2024) reformulates next POI recommendation as a text generation task by serializing check-in trajectories into natural language prompts. It leverages the pre-trained knowledge of LLMs to understand location semantics through rich textual descriptions of POI names and categories.

GNPR-SID. GNPR-SID (Wang et al., 2025) introduces discrete semantic identifiers to represent POIs, replacing verbose POI names with compact token sequences derived from clustering POI embeddings. This reduces input length and enables compositional generalization across POIs with shared semantic properties.

GA-LLM. GA-LLM (Liu et al., 2025) injects geographic coordinates and POI-to-POI transition structures into the LLM-based recommendation framework. It encodes latitude-longitude information and spatial distance features directly into the prompt, enhancing the model's awareness of geographic constraints in user mobility.

CoAST. CoAST (Zhai et al., 2025) introduces semantic discrete identifiers combined with a cognitive alignment training strategy. It designs a multi-stage training pipeline that progressively aligns the LLM's representations with spatial-temporal patterns in user trajectories.

ROS. ROS (Lv et al., 2026) represents the current state-of-the-art in generative POI recommendation. It introduces Hierarchical Spatial Semantic IDs that encode both geographic locality (via S2 cell prefixes) and semantic similarity into compositional token sequences. ROS further proposes a Mobility Chain-of-Thought reasoning paradigm and spatial-guided reinforcement learning to enhance prediction accuracy. AWARE builds upon the same SID-equipped backbone as ROS, enabling a controlled comparison that isolates the contribution of world knowledge injection.

E Implementation Details

Implementation Details. AWARE adopts `Qwen3-4B` (Yang et al., 2025) as the generative backbone, with Spatial Semantic IDs derived from `Qwen3-0.6B` embeddings. To preserve the pre-trained spatial reasoning capability while adapting to the knowledge-enriched input, we apply LoRA (Hu et al., 2022) for parameter-efficient fine-tuning with rank $r = 64$ and scaling factor $\alpha = 128$. The world knowledge is generated by prompting an LLM agent (`Qwen3-max`) with each user's trajectory and temporal context prior to training. The agent retrieves external evidence via SerpAPI-backed web search with date-range filtering and page-level content extraction, capped at two parallel retrieval rounds per user. Output length is regulated by a rewrite-then-truncate cascade. For time-aware history selection, we set the periodic ratio $\beta = 0.4$ and the history budget $L = 150$ entries. Training is conducted for 2 epochs with a learning rate of 1e-4 on $8 \times$ AMD-Instinct-MI308X-OAM GPUs. For a fair comparison, we inherit all other hyperparameters from ROS (Lv et al., 2026).

Temporal Fidelity. The evaluation benchmarks (Foursquare NYC/TKY, Gowalla-CA) span 2009–2013 and predate the training corpora of contemporary LLMs, risking contamination of historical-user narratives with present-day information. To prevent this, temporal grounding is enforced at three levels. *(i) Prompt level:* Π instructs the agent to treat the latest timestamp in $\mathcal{S}_u$ as "now" and to prefer archival sources over undated present-day pages. *(ii) Tool level:* every search query is rewritten to inject the target year and month as a date-range constraint, and results are pre-filtered by publication date. *(iii) Synthesis level:* a post-hoc verifier discards any claim not corroborated by a source dated within $[t - \Delta,\, t + \Delta]$ ($\Delta = 30$ days). Under this regime, the injected world knowledge faithfully reflects the user's original temporal and spatial context. A manual audit of 200 randomly sampled narratives confirms that event references are anchored within the target window.

F Case Study

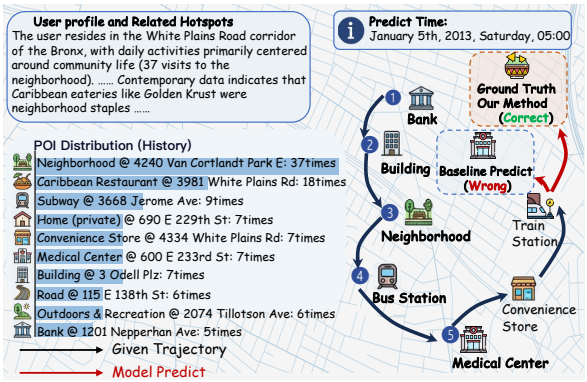

Figure 7: Case study on NYC. The user frequents a neighborhood area (37 visits) and a Caribbean restaurant (18 visits). ROS predicts *Train Station* from the recent trajectory, while AWARE correctly predicts *Caribbean Restaurant*.

To illustrate how world knowledge captures intent beyond sequential patterns, we present a representative case in Figure 7. On Saturday morning, ROS predicts *Train Station* based on the recent trajectory ending near a transit hub. AWARE instead correctly predicts *Caribbean Restaurant*, guided by the following world knowledge:

> *"Contemporary data indicates that **Caribbean eateries like Golden Krust were neighborhood staples** ... With cold weather prevailing, the user is likely seeking convenient, familiar services nearby, particularly **warm Caribbean meals**."*

Behavioral priors further corroborate the prediction: POI frequency statistics rank Caribbean Restaurant as the user's second most visited location, and transition patterns reveal 13 bidirectional trips between it and the nearby neighborhood area. This case illustrates how world knowledge provides the cultural *why* behind a visit, while behavioral priors supply the personalized *where*, together resolving predictions that sequential models cannot.

Instruction:
Here is a record of a user's POI accesses, your task is based on the history to predict the POI that the user is likely to access at the specified time.

Input:
<user_poi_stats>
User frequently visits:
Coffee Shop at <m_152><n_175><a_0><b_22><c_0> (2 times),
American Restaurant at <m_136><n_88><a_11><b_13><c_0> (2 times),
Bar at <m_132><n_223><a_10><b_5><c_0> (1 times),
Bar at <m_133><n_104><a_26><b_5><c_0> (1 times),
Hotel at <m_143><n_207><a_5><b_2><c_0> (1 times),
Bar at <m_142><n_55><a_26><b_17><c_0> (1 times),
Burger Joint at <m_142><n_72><a_16><b_11><c_0> (1 times),
American Restaurant at <m_155><n_247><a_0><b_13><c_0> (1 times),
Diner at <m_153><n_74><a_0><b_14><c_0> (1 times),
French Restaurant at <m_139><n_178><a_0><b_25><c_0> (1 times).
</user_poi_stats>

<transition_patterns>
User transition patterns:
<m_132><n_223><a_10><b_5><c_0> → <m_133><n_104><a_26><b_5><c_0> (1 times),
<m_133><n_104><a_26><b_5><c_0> → <m_143><n_207><a_5><b_2><c_0> (1 times),
<m_143><n_207><a_5><b_2><c_0> → <m_142><n_55><a_26><b_17><c_0> (1 times),
<m_142><n_55><a_26><b_17><c_0> → <m_142><n_72><a_16><b_11><c_0> (1 times),
<m_152><n_175><a_0><b_22><c_0> → <m_155><n_247><a_0><b_13><c_0> (1 times),
<m_155><n_247><a_0><b_13><c_0> → <m_153><n_74><a_0><b_14><c_0> (1 times),
<m_153><n_74><a_0><b_14><c_0> → <m_139><n_178><a_0><b_25><c_0> (1 times),
<m_139><n_178><a_0><b_25><c_0> → <m_152><n_81><a_0><b_13><c_0> (1 times),
<m_152><n_81><a_0><b_13><c_0> → <m_153><n_26><a_26><b_17><c_0> (1 times),
<m_153><n_26><a_26><b_17><c_0> → <m_136><n_88><a_11><b_13><c_0> (1 times).
</transition_patterns>

<user_preference>
The user's trajectory in mid-April 2012 shows a pattern of visiting bars across Midtown Manhattan, suggesting an interest in nightlife and cocktail culture during a short stay possibly centered around leisure or business near Park Avenue. At that time, NYC's bar scene was dominated by craft cocktail venues like Death & Company and PDT, while Midtown hosted live music events and emerging indie performances. Given this context and the user's repeated late-night bar visits, they were likely seeking trendy, high-quality cocktail bars or live music venues in central Manhattan for evening entertainment.
</user_preference>
Given user behavior sequence:
April 13th, 2012, Friday, 09:11, visit Bar at 599 10th Ave <m_132><n_223><a_10><b_5><c_0>.
April 13th, 2012, Friday, 12:45, visit Bar at 235 W 48th St <m_133><n_104><a_26><b_5><c_0>, distance is Nearby.
April 14th, 2012, Saturday, 01:45, visit Hotel at 301 Park Ave <m_143><n_207><a_5><b_2><c_0>, distance is Nearby.
April 14th, 2012, Saturday, 06:07, visit Bar at 218 E 53rd St <m_142><n_55><a_26><b_17><c_0>, distance is Nearby.
At April 14th, 2012, Saturday, 07:15, user will visit

Output: <m_142><n_72><a_16><b_11><c_0>

Table 6: A complete AWARE prompt example from NYC. The <user_preference> block (highlighted) contains the world knowledge injected by AWARE, which is the key difference from the ROS baseline.

Role and Input You are an assistant designed to support a POI recommendation system for users in `{city}`. Your input is a user's historical POI visit sequence in chronological order. Each line is one past visit record with time, POI ID, POI type, address, latitude, and longitude.
Reasoning Protocol Based on the user's historical POI visits and the current date implied by the latest part of the trajectory, you should: (1) infer the user's likely preferences, routines, and near-term intent; (2) infer the most relevant current area or activity context from the user's recent visits; (3) use web tools to search for social hotspots, local events, lifestyle trends, seasonal patterns, and timely popular activities that are relevant to `{city}`, the inferred area, and the current date; (4) produce a compact summary that connects user history with current trends and predicts what kinds of POIs the user is most likely to be interested in next.
Tool Usage Requirements • Actively use `web_search` to gather timely signals from the public web. • Use `web_fetch` to verify high-value pages when needed. • Do not rely on a single search result or a single source. • Prefer signals that are timely, local, behaviorally relevant, and useful for POI recommendation.
Reasoning Focus • Use the historical POI sequence to infer recurring interests such as dining style, commute pattern, nightlife preference, fitness habits, shopping preference, tourism behavior, or work-related mobility. • Pay more attention to the user's most recent visits when inferring current intent. • Combine user preference signals with timely external signals such as local events, weather-related behavior shifts, seasonal demand, holidays, viral venues, neighborhood buzz, and popular urban activities. • Make a grounded prediction about the next POI categories, areas, or venue styles the user may want.
Output Constraints • Output only one short paragraph in English. • No JSON, Markdown, bullet points, titles, or explanations. • Keep the text compact, information-dense, and directly useful for recommendation. • The paragraph must tightly integrate three elements: what the user's historical behavior suggests, what current hotspots or trends are relevant, and what POIs the user is likely to seek next. • Avoid generic wording and trajectory recitation. • Maximum length: `{max_words}` words.

Table 7: System prompt for the world knowledge acquisition agent. The prompt is instantiated for the NYC dataset with a 150-word budget.