BEYOND ROUTINE: A CONTROLLED BENCHMARK FOR EVENT GROUNDING IN NEXT-POI PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Next-point-of-interest (next-POI) prediction forecasts a user's next destination from their observed visit history. Events provide additional context, but predictive event utility does not necessarily imply event grounding. Context-on/off evaluation can reveal a gain without establishing whether it depends on the supported temporal and spatial bindings. We introduce BEYONDROUTINE, a controlled benchmark for event grounding in next-POI prediction under an event-aware input setting. Its construction combines chronological mobility records with independently documented events and requires evidence that event information was available before prediction. Events are matched to prediction instances without consulting the future target. Each comparison fixes the prediction instance, candidates, target, and trained model while removing event information or changing its time or location. We further introduce EventResidual, a reference predictor that factorizes history and event representations and applies gated, candidate-specific score corrections over shared POI representations. In the event-conditioned model study, the evaluated predictors show different utility and binding profiles, while stricter controls reveal additional variation in binding sensitivity. EventResidual shows positive paired point estimates on all three primary evaluation sets, including an event-utility difference of 0.0233 in MRR@10.

1 INTRODUCTION

Next-point-of-interest (next-POI) prediction estimates which place a user will visit next from their time-ordered visit history (Luca et al., 2023a). Established methods learn personal preferences, spatial transitions, and temporal regularities from check-in data (Feng et al., 2018; Yang et al., 2020; Luo et al., 2021; Yang et al., 2022; Xu et al., 2024). These signals capture recurring mobility patterns, but do not directly describe new activities occurring in a city. A scheduled local event may make a destination relevant even when it is absent from a user's routine. This motivates studying how event information available at prediction time can complement historical evidence for next-POI prediction.

Recent work provides foundations for this setting. AWARE incorporates dynamic world knowledge into next-POI recommendation, ELLMob contributes event-annotated data and trajectory generation, and MSAHG models scenario-dependent mobility (Ding et al., 2026; Wang et al., 2026; Lin et al., 2026). Here, next-POI prediction is the task, event-aware describes the added input, and event grounding is the scientific question: does the predictor use the event's supported time and place? *Predictive event utility does not necessarily imply event grounding.* Familiar venue cues, for example, may remain useful after an event's time is changed.

Research questions. **RQ1:** How well can event-associated prediction instances be balanced against a routine reference using pre-cutoff information? **RQ2:** When event input helps prediction, does performance also depend on its supported temporal and spatial bindings? **RQ3:** How stable are EventResidual's utility and binding effects across unfamiliar episodes and stricter controls?

Answering these questions across models requires a common evaluation, not only method-specific input ablations. We introduce BEYONDROUTINE, whose construction links mobility records, documented events, and paired predictions. Separately documented events are associated with prediction instances through time, spatial bindings, and evidence of information availability. Event selection excludes the future target. The benchmark covers sports, weather alerts, entertainment, and sup-

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

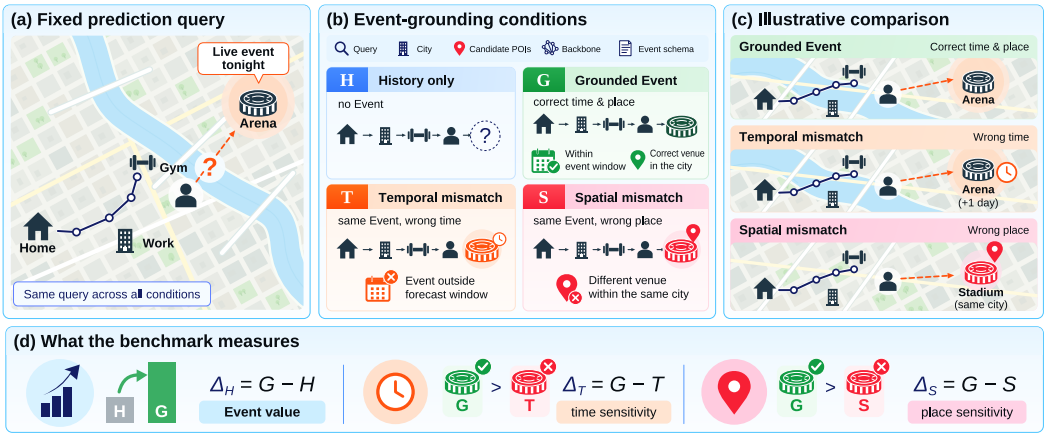

Figure 1: **Event-input comparisons on a fixed prediction instance.** The history, candidates, and observed target remain fixed while the selected event is removed (H), retained (G), given a time substitution (T), or given a supported same-city spatial substitution (S). The point-venue diagram depicts the comparison protocol rather than quantitative outcomes. Eligibility is defined in Section 3.3 and Appendix B.1.

plementary civic records. Temporal and spatial comparisons use their own constructible subsets (Table 2).

The benchmark compares each selected event with no event information, a time-shifted version, or a version moved to another documented event location in the same city (Figure 1). Each pair retains the history, candidates, target, and trained predictor. Event removal measures context utility; time/place substitutions test binding dependence. Matching diagnostics describe the comparability of an event-associated population and a routine reference; they are separate from the within-instance input comparisons.

We also introduce **EventResidual**, a reference predictor that learns separate history and event representations over shared candidate POI representations. The history branch supplies an initial ranking, and a gate scales the event branch's candidate-specific score corrections. Our contributions are an event-grounded benchmark built from independently sourced evidence with prediction-time admissibility; target-independent fixed-instance comparisons that distinguish context utility from binding dependence; and EventResidual, a factorized reference predictor with gated candidate-wise corrections. Sections 5.2–5.4 follow RQ1–RQ3 in that order.

In the model study, EventResidual shows positive paired point estimates on the primary evaluation sets, whereas the evaluated ELLMob adaptation has negative estimates. EventResidual retains positive utility estimates on unseen episodes, while stricter controls reveal additional variation in temporal and spatial binding sensitivity (Section 5.4). The results support measuring utility and binding separately and reporting their dependence on the evaluation population.

2 Related Work

2.1 Mobility Datasets and Benchmarks

Mobility benchmarks study complementary aspects of prediction. YJMob100K covers routine and emergency periods, HuMob'24 evaluates multi-city prediction, and Massive-STEPS combines semantic check-ins with supervised, zero-shot, and spatiotemporal reasoning evaluations (Yabe et al., 2024b;a; Wongso et al., 2026). CityBench evaluates urban tasks, while USTBench examines reasoning processes as well as task outcomes (Feng et al., 2025b; Lai et al., 2026). Trajectory-overlap analysis further motivates separating generalization from train–test familiarity (Luca et al., 2023b). Table 1 positions the directly related datasets and benchmarks by task, context linkage, and evaluation focus.

Table 1: **Dataset and benchmark positioning.** Linkage granularity describes how context is organized for evaluation, not the resolution of recorded trajectories.

Benchmark	Task	External context	Linkage granularity	Evaluation focus
YJMob100K	Grid-location prediction	POI category counts; period regimes	Period-level regime; grid-cell POI context	Routine versus emergency predictability.
HuMob'24	Multi-city trajectory prediction	Grid-cell POI category counts	City-level transfer; grid-cell context	Cross-city transfer and trajectory accuracy.
Massive-STEPS	POI prediction; trajectory reasoning	Semantic POI metadata	POI / trajectory-level	Semantic, geographic, and zero-shot generalization.
BEYONDROUTINE	Event-aware next-POI prediction	Independently sourced event evidence	Prediction-instance-level event linkage	Event utility and temporal / spatial binding.

Why reconstruction is needed. Existing mobility datasets do not generally provide cutoff-time event evidence, target-independent event assignment, and time/place comparisons on a fixed prediction instance together. The missing combination leaves an identification gap: context-on/off gains alone do not establish dependence on supported event bindings. BEYONDROUTINE constructs these linked records for that comparison.

2.2 NEXT-POI PREDICTION METHODS

Historical next-POI predictors learn preferences and sequential, spatial, or temporal regularities. ST-RNN explicitly models time intervals and geographical distances between observed visits (Liu et al., 2016). DeepMove, Flashback, and STAN model dependencies in past visits; GETNext incorporates trajectory-flow structure; and LoTNext addresses long-tailed prediction (Feng et al., 2018; Yang et al., 2020; Luo et al., 2021; Yang et al., 2022; Xu et al., 2024). GeoMamba and MSAHG extend geographical and scenario-specific representations (Qin et al., 2025; Lin et al., 2026). They provide the historical and contextual prediction methods against which event information must be assessed.

Context-enhanced methods extend the available information or its representation. AWARE incorporates dynamic world knowledge, LLM4POI uses heterogeneous trajectory context, and ROS studies geographic reasoning for generative next-POI recommendation (Ding et al., 2026; Li et al., 2024; Lv et al., 2026). ELLMob instead addresses event-conditioned trajectory generation, so using it for single-step prediction requires an explicit task adaptation (Wang et al., 2026). AWARE uses date-filtered retrieval and component ablations; ELLMob supplies common event context to its baselines. These evaluations assess context utility and generation fidelity. BEYONDROUTINE instead pairs supported and substituted bindings for the same next-POI instance and fitted predictor; EventResidual provides a factorized reference predictor within that protocol.

3 BEYONDROUTINE

Three principles organize construction: *independently sourced event evidence*, *prediction-time admissibility*, and *target-independent fixed-instance comparison*. Mobility records and event sources are acquired separately, rather than finding anomalies and retrospectively assigning an event explanation. Supported time/place information constrains context eligibility; target-independent assignment precedes closed-set filtering and input substitutions. Together these principles supply the evidence and paired inputs needed to identify binding dependence, beyond context-on/off prediction scores.

3.1 MOBILITY RECORDS AND PREDICTION TASK

Following the sequential next-POI formulation in Flashback and GETNext (Yang et al., 2020; 2022), an observed visit prefix predicts the immediately following recorded POI. For user u, let

$H_q = ((p_1, t_1), \ldots, (p_i, t_i))$, with prediction cutoff $t_q = t_i$ and target $y_q = p_{i+1}$. We call this example a *prediction instance*, indexed by q. The event-aware extension adds context E_q available by t_q; it does not supply the future target timestamp.

BEYONDROUTINE uses a training-known, city-conditioned candidate vocabulary as its evaluation policy. Chronological training records determine the user's most frequently visited city c_u and the POIs $\mathcal{L}_{c_u}$ observed there during training. Earlier visits from other cities remain in H_q. The prediction task is

$$C_q = \mathcal{L}_{c_u}, \qquad \hat{y}_q = \arg\max_{\ell \in C_q} p_\theta(\ell \mid H_q, E_q). \tag{1}$$

Candidates are fixed before validation or test targets are inspected. *Closed set* requires vocabulary membership, not a prior visit by this user: personally unseen but training-known POIs remain eligible. New-vocabulary and cross-city targets are outside this evaluation. Appendix B specifies this candidate construction.

3.2 EVENT COLLECTION AND PREDICTION-TIME EVIDENCE

Events are collected without consulting next-visit targets. The inventory uses six source/reporting categories: baseball, other scheduled sports, severe-weather warnings, severe-weather watches, entertainment mass gatherings, and civic public gatherings. The principal families are scheduled sports, weather alerts, and entertainment; civic records provide supplementary coverage (Appendix A). An *episode* is a benchmark event record with semantic fields, occurrence-time information, a spatial binding, and source references. Counts use episode IDs rather than cross-source-deduplicated physical events; one ID may support many prediction instances. Source provenance is retained separately from model inputs (Appendix A), following the distinction between dataset composition, collection, and intended use emphasized by Gebru et al. (2021).

We distinguish *when an event occurs* from *when its information becomes available*. Let a_e be the supported availability time of the fields supplied to the model. They are eligible only if $a_e \leq t_q$; an archive capture supports availability by that capture, not necessarily the original publication time. Date-precision evidence remains `DATEONLY`; a time discovered later cannot upgrade it to `DATETIME`. Availability evidence, retrieval intervals, and occurrence-time bindings remain separate.

Event frequency and prediction-instance volume describe complementary aspects of temporal coverage (Figure 2). Appendix A adds weekly activity profiles.

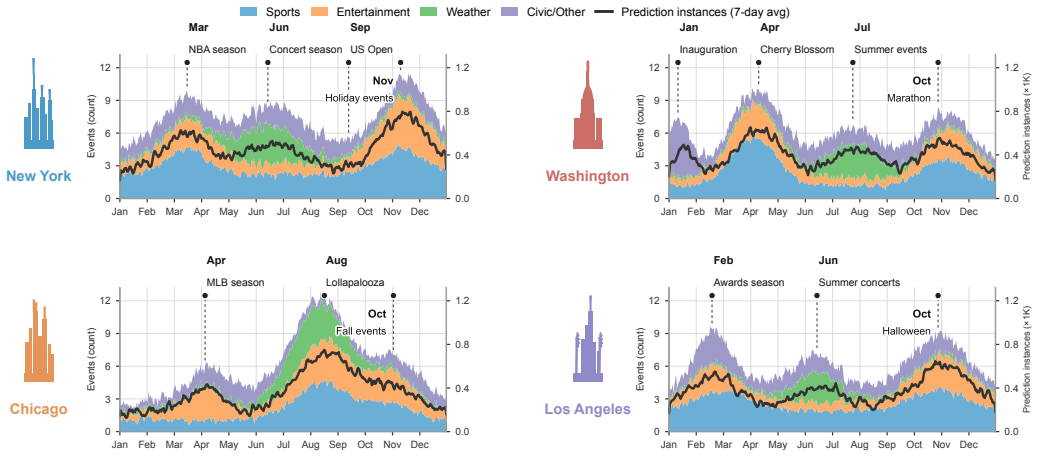

Figure 2: **City–event rhythms.** Stacked areas show event-frequency profiles by category; the line shows seven-day-average prediction-instance activity. Event frequency and prediction-instance activity use separate vertical axes.

3.3 Event Assignment and Controlled Inputs

For event e, let $\mathcal{C}_e$ be its documented city bindings (one city for a point venue, potentially several for a regional event) and I_e^{ret} its retrieval interval. Eligibility requires

$$\mathcal{A}(q) = \left\{ e : \; a_e \le t_q, \; c_u \in \mathcal{C}_e, \; I_e^{\text{ret}} \cap [t_q, t_q + 24\text{h}) \neq \varnothing \right\}. \tag{2}$$

Spatial grounding concerns an event's physical influence domain, which may extend well beyond a venue and need not follow a municipal boundary. A supported binding is the event record's documented time or place; its spatial component is a venue, route, or affected-area geometry. Equation 2 uses the recorded event-to-city mapping for coarse retrieval of this binding; it does not estimate the influence boundary. The 24-hour window limits retrieval, not the next-visit time. Eligible events are ordered by temporal gap, then episode identifier. Filtering on future target distance would leak the outcome into assignment. G/S tests dependence on the supplied place binding. Appendix B.1 summarizes assignment. Let Q_{event} denote the selected, event-associated instances in a named split. Closed-set evaluation retains exactly

$$Q_{\text{closed}} = \{ q \in Q_{\text{event}} : y_q \in C_q \}. \tag{3}$$

This target-membership check is applied *after* event assignment: it does not choose the reference city, candidate set, or selected event.

Let s_e, b_e^t, b_e^s be the selected event's semantic content, time binding, and place binding. After applying any comparison-specific identity mask, the four inputs are

$$\begin{aligned} E_q^H &= \varnothing, & E_q^G &= \{(s_e, b_e^t, b_e^s)\}, \\ E_q^T &= \{(s_e, \mathcal{T}_q(b_e^t), b_e^s)\}, & E_q^S &= \{(s_e, b_e^t, \mathcal{S}_q(b_e^s))\}. \end{aligned} \tag{4}$$

H removes the event and G retains it. T changes the supported occurrence-time binding while preserving place and non-temporal content; valid exact-time instances form Q_T. S substitutes a family-compatible, same-city spatial binding while preserving time and non-spatial content; valid instances form Q_S. Where required, identifying names are masked in both spatial-comparison arms, while the explicit binding under test is retained. Figure 1 illustrates point venues; family-specific geometries and validity checks appear in Appendix B.1.

Controls operate after selection: the altered event is neither reselected nor removed by the original retrieval filter. Binding-derived features change with the binding; history, candidates, observed target, and fitted predictor remain fixed. Q_T and Q_S are separate constructible subsets, not successive filtering stages.

3.4 Quality Checks and Benchmark Coverage

Table 2: **Corpus and closed-test scope.** Multi-category benchmark counts; controls may overlap.

Mobility corpus	
Prediction instances	27.5M
Users	185,477
Closed-test event population	
Event-associated instances	40,221
Users	5,641
Selected episode IDs	1,315
Source/reporting categories	6
Control populations	
Temporal controls $\lvert Q_T \rvert$	9,649
Spatial controls $\lvert Q_S \rvert$	10,697

Construction checks cover split roles, availability evidence, event assignment, target membership, time/coordinate interpretation, and paired-field consistency. Predictive gains are not an admission criterion. Table 2 distinguishes the source mobility corpus from the multi-category event-grounding evaluation population. Each count identifies a different unit: a prediction instance, a user, or a selected episode ID. One episode may support many instances, so episode support is reported separately from sample volume.

The closed-test inventory contains 40,221 event-associated instances and 1,315 selected episode IDs across six source/reporting categories. Time and place comparisons use separate constructible subsets rather than successive filtering stages. Section 5 evaluates the event-study subset with its own paired populations, while Table 2 reports the broader multi-category benchmark inventory. Appendix A supplies all-split counts and time precision, while Appendix C.1 gives the model-study populations.

Coverage is measured in both directions (Figure 3). *Instance coverage* is the fraction of candidate instances with an eligible event; *event reach* is the fraction of admitted event–city bindings reached by those instances. The displayed cities have different combinations of these rates and user support. Many covered instances can rely on recurring events, while a large catalogue may overlap relatively few recorded visits. Neither rate measures target-vocabulary retention or predictive utility. Appendix A gives counting units and additional category and temporal profiles.

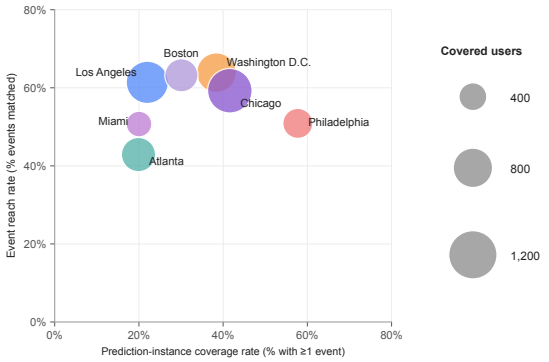

Figure 3: **Two-way coverage.** Instance coverage and event reach; bubble area denotes distinct covered users. Counting units: Appendix A.

4 EventResidual

EventResidual learns how event context adjusts a history-based next-POI ranking. Separate pathways encode routine mobility and supplied events, while both score the same candidate POIs (Figure 4). The experiments assess the predictor's event utility, binding dependence, and behavior across episode-familiarity and stricter-control subsets.

4.1 Representations and Gated Ranking

Let $\mathbf{h}_q = f_H(H_q)$ summarize recent and longer-term history, $\mathbf{r}_q = f_E(E_q)$ encode the event's semantics and explicit time/place inputs, and $\mathbf{v}_\ell = f_P(\ell)$ represent candidate ℓ. The history score is

$$z^{\text{hist}}_{q,\ell} = s_H(\mathbf{h}_q, \mathbf{v}_\ell), \qquad \ell \in C_q. \tag{5}$$

The event pathway computes a raw candidate residual and a scalar gate,

$$\delta_{q,\ell} = s_E(\mathbf{h}_q, \mathbf{r}_q, \mathbf{v}_\ell), \qquad \alpha_q = \mathbf{1}\{E_q \neq \varnothing\}\, \sigma(g(\mathbf{h}_q, \mathbf{r}_q))\,, \tag{6}$$

which yield

$$c_{q,\ell} = \alpha_q \delta_{q,\ell}, \qquad z_{q,\ell} = z^{\text{hist}}_{q,\ell} + c_{q,\ell}. \tag{7}$$

Candidate-specific corrections can alter relative scores; a common additive offset cannot. The gate scales the update and explicitly disables it for empty event input. It is not an exposure-probability estimate. The diagram's $z_H, \Delta_E, g(E)$ denote $z^{\text{hist}}, \delta, \alpha$ in these equations.

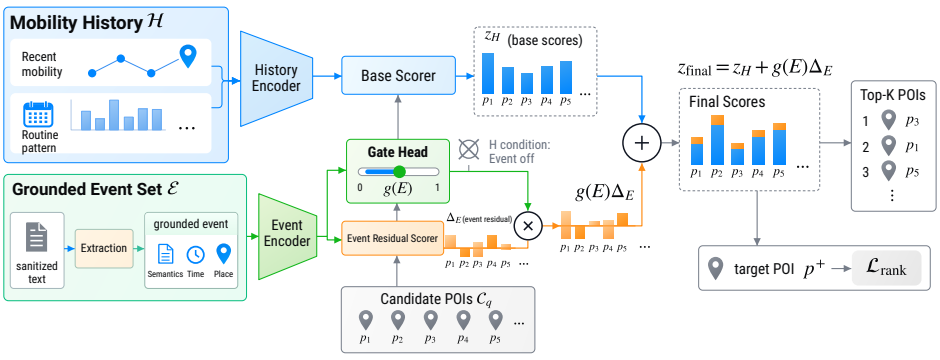

Figure 4: **EventResidual architecture.** History and grounded-event pathways score the same candidate POIs. A scalar gate modulates candidate-wise event residuals; the target enters only the ranking loss. The diagram's z_H, Δ_E, and $g(E)$ correspond to the base scores, raw residuals, and gate in Equations 5–7. Time and spatial substitutions are evaluation inputs.

4.2 Training and Implementation

EventResidual is implemented in PyTorch with separate history and event encoders and shared candidate-POI representations. The history encoder combines the 20 most recent visits with a longer-horizon summary over up to 60 active days and four visits per day; its recent-history Transformer uses two layers and four attention heads. The event encoder uses two multi-head-attention layers with four heads and combines semantic event features with explicit temporal, spatial, quality, and event-family inputs. Candidate-wise event corrections are bounded by a residual cap of 0.5 and scaled by a learned scalar gate; empty event input disables the correction, so H is the event-off forward pass of the same fitted model rather than a separately trained history model. Training uses AdamW with learning rate 3×10^{-4}, weight decay 10^{-2}, and gradient-norm clipping at 1.0. The objective jointly optimizes grounded-event and event-masked ranking paths, while time and location substitutions are used only for evaluation. Model selection uses validation MRR@10 on the grounded-event arm, and Appendix B.1 reports additional implementation and reproducibility details.

5 Experiments

We compare EventResidual and an ELLMob next-POI adaptation through event removal, time/place substitutions, and episode-level diagnostic slices.

5.1 Experimental Setup

Evaluation scope. The model study evaluates a collection of 291 documented event episodes. Here Q_{closed}, Q_T, and Q_S contain 18,055, 5,761, and 11,094 instances, respectively (full support in Appendix C.1). These are the model-study populations, distinct from the multi-category inventory in Section 3. Spatial substitution is supported in New York and Chicago. Candidates follow the training-known, closed-set policy in Section 3.1.

Models and conditions. ELLMob denotes the evaluated next-POI adaptation (Wang et al., 2026). H removes event context from the fitted predictor; G supplies the selected event; T and S change its time and place bindings, respectively. Paired conditions retain the history, candidate set, target, and fitted model. H is not a separately trained history baseline.

Metrics. MRR@10 is primary; Acc@1/5/10 are supplementary. For per-instance score m_q, we use the paired performance effects

$$d_q^{A,B} = m_q^A - m_q^B, \qquad \Delta^{A,B}(Q) = |Q|^{-1} \sum_{q \in Q} d_q^{A,B}, \tag{8}$$

$$\Delta_H = \Delta^{G,H}(Q_{\text{closed}}), \quad \Delta_T = \Delta^{G,T}(Q_T), \quad \Delta_S = \Delta^{G,S}(Q_S). \tag{9}$$

Δ_H measures event utility; Δ_T and Δ_S measure binding sensitivity. Main results are instance-weighted point estimates, with each pair evaluated on its own population. Appendix C gives full population counts, accuracy, equal-episode estimates, and available clustered intervals.

5.2 RQ1: Comparability with a Routine Reference

Pre-cutoff matching leaves 9.94% of instances unmatched, with no routine-reference reuse. Standardized mean differences range from 0.028 to 0.052 across five covariate groups (Appendix C.1), supporting comparability on the measured covariates (Austin, 2009). These diagnostics establish balance on the measured pre-cutoff covariates; model-level event effects are assessed separately through the fixed-instance H/G comparison below.

5.3 RQ2: Event Utility and Binding Dependence

EventResidual shows positive paired effects on the primary sets. Table 3 and Figure 5 show positive point estimates for event utility ($+0.0233$), temporal binding ($+0.0354$), and spatial binding ($+0.0305$). On Q_{closed}, Acc@1 is 22.28% under G and 19.09% under H.

Table 3: **Event utility and time/place binding effects.** Positive paired effects favor G; each block uses its own eligible population. Values are MRR@10 point estimates.

Model	G	Control	Control MRR	Paired effect
Event utility: $G - H$ *on* Q_{closed}*;* $N = 18,055$.				
EventResidual	0.2809	H	0.2576	**+0.0233**
ELLMob adaptation	0.2216	H	0.2436	-0.0220
Temporal binding: $G - T$ *on* Q_T*;* $N = 5,761$.				
EventResidual	0.2977	T	0.2623	**+0.0354**
ELLMob adaptation	0.2243	T	0.2320	-0.0077
Spatial binding: $G - S$ *on* Q_S*;* $N = 11,094$.				
EventResidual	0.2778	S	0.2473	**+0.0305**
ELLMob adaptation	0.2188	S	0.2253	-0.0065

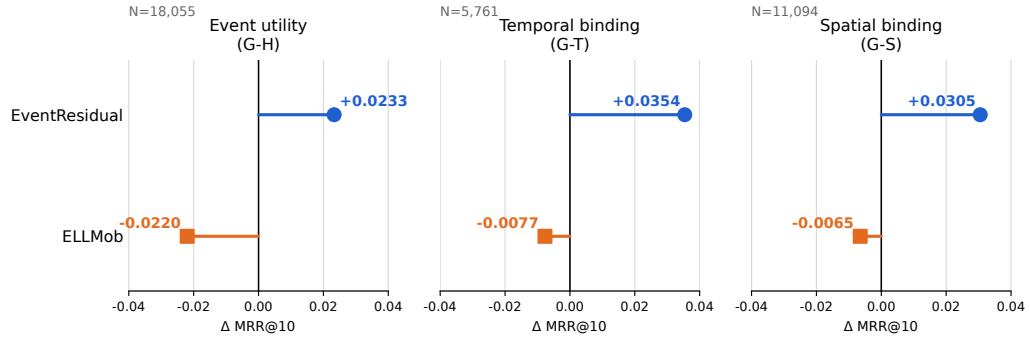

Figure 5: **Grounding diagnostic profiles.** Paired MRR@10 point estimates for event utility ($G - H$), temporal binding ($G - T$), and spatial binding ($G - S$). Positive values favor G, and each panel uses its own eligible population shared by both evaluated models.

The ELLMob adaptation responds differently. Its event utility is negative (-0.0220), including on the exact-time subset (-0.0151). The reported user- and episode-cluster intervals for both utility estimates lie below zero (Appendix C.2). Its temporal and spatial effects are also negative. These findings characterize the evaluated ELLMob next-POI adaptation under the BEYONDROUTINE protocol.

What the paired evaluation adds. Context-on/off scores measure event utility, whereas G/T and G/S measure sensitivity to the supplied bindings. The distinct model profiles motivate reporting all three effects rather than treating event utility as a substitute for binding tests.

5.4 RQ3: ROBUSTNESS ACROSS DIAGNOSTIC SUBSETS

Table 4: **EventResidual across subsets.** Paired MRR@10 differences.

Subset	Instances	Effect
Unseen ($G - H$)	1,384	+0.0130
Low similarity ($G - H$)	656	+0.0111
Strict temporal ($G - T$)	3,138	-0.0067
Hard spatial ($G - S$)	4,816	+0.0022

EventResidual retains positive event-utility estimates on unseen and low-similarity episodes (Table 4). Equal-episode effects are $+0.0055$ in both cases, so the sign persists when frequently represented episodes no longer dominate the average.

Binding estimates are less uniform. The strict temporal effect is -0.0067, opposite in sign to the primary estimate; the hard spatial effect remains positive at $+0.0022$. Equal-episode estimates retain these signs (Appendix C.2). The contrast between primary and stricter subsets shows that observed binding dependence is sensitive to the evaluation population.

6 Discussion, Limitations, and Conclusion

The primary point estimates differ across models: EventResidual has positive utility and binding effects, while the evaluated ELLMob adaptation has negative effects. Stricter subsets reveal population-dependent variation in binding sensitivity. These results support measuring event utility and binding dependence separately and specifying the population for each estimate.

The model study evaluates the designated event-study populations, with spatial controls supported on the eligible New York and Chicago instances. Matching is used to establish covariate comparability for the routine reference, while the within-instance $H/G/T/S$ comparisons provide the model-level diagnostics. EventResidual results are therefore interpreted as paired point estimates; attribution to a specific integration mechanism would require a matched architectural comparison.

The interventions diagnose model dependence on supplied event bindings rather than causal effects of events on human mobility. The results show why event utility and dependence on supported bindings should be reported separately rather than inferred from a single event-aware score.

AI Use Statement

Generative AI tools, including ChatGPT and AI coding assistants, supported conceptual framing and research questions, evaluation-design review, literature organization and source lookup, interpretation of author-supplied experimental summaries, manuscript drafting and editing, LaTeX and tables, data-processing and plotting code, and figure development. Experimental measurements were produced and verified by the research team. The effect plot was rendered from those measurements. AI-assisted checks covered arithmetic, population consistency, references, and rendered-page inspection. The authors retain responsibility for final verification, scientific review, and submission-form disclosure.

Ethics Statement

Human mobility records carry privacy and re-identification risks. Release materials will document pseudonymization, geographic/temporal resolution, source licenses, and withheld fields. Event linkage does not establish individual intent or participation.

Reproducibility Statement

Sections 3 and 5.1 specify the task, input conditions, population definitions, and paired metrics. Appendix B.1 distinguishes code-confirmed source settings from run-specific configuration fields. Appendix C reports supporting scores, population counts, aggregation variants, and available clustered intervals, while Appendix A summarizes data sources and provenance. Access to source records remains subject to their licenses.

References

Peter C. Austin. Balance Diagnostics for Comparing the Distribution of Baseline Covariates between Treatment Groups in Propensity-Score Matched Samples. *Statistics in Medicine, 28(25):3083–3107*, 2009.

Qiuyu Ding, Heng-Da Xu, Wei Zhang, Dongyi Lv, Changda Xia, Feng Xiong, and Mu Xu. Why Users Go There: World Knowledge-Augmented Generative Next POI Recommendation. *arXiv:2605.11807v1*, 2026.

Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. DeepMove: Predicting Human Mobility with Attentional Recurrent Networks. *WWW, pp. 1459–1468*, 2018.

Jie Feng, Yuwei Du, Jie Zhao, and Yong Li. AgentMove: A Large Language Model based Agentic Framework for Zero-shot Next Location Prediction. *NAACL, pp. 1322–1338*, 2025a. doi:10.18653/v1/2025.naacl-long.61.

Jie Feng, Jun Zhang, Tianhui Liu, Xin Zhang, Tianjian Ouyang, Junbo Yan, Yuwei Du, Siqi Guo, and Yong Li. CityBench: Evaluating the Capabilities of Large Language Models for Urban Tasks. *KDD*, 2025b.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *Communications of the ACM*, 2021.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-Based Recommendations with Recurrent Neural Networks. *ICLR*, 2016.

Siqi Lai, Yansong Ning, Zirui Yuan, Zhixi Chen, and Hao Liu. USTBench: Benchmarking and Dissecting Spatiotemporal Reasoning Capabilities of LLMs as Urban Agents. *ICLR*, 2026.

Peibo Li, Maarten de Rijke, Hao Xue, Shuang Ao, Yang Song, and Flora D. Salim. Large Language Models for Next Point-of-Interest Recommendation. *SIGIR, pp. 1463–1472*, 2024.

Yuxi Lin, Yongkang Li, Jie Xing, and Zipei Fan. Multifaceted Scenario-Aware Hypergraph Learning for Next POI Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence, 40(18):15269–15277*, 2026.

Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. *AAAI, pp. 194–200*, 2016.

Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. A Survey on Deep Learning for Human Mobility. *ACM Computing Surveys, 55(1):1–44*, 2023a.

Massimiliano Luca, Luca Pappalardo, Bruno Lepri, and Gianni Barlacchi. Trajectory Test–Train Overlap in Next-Location Prediction Datasets. *Machine Learning, 112:4597–4634*, 2023b.

Yingtao Luo, Qiang Liu, and Zhaocheng Liu. STAN: Spatio-Temporal Attention Network for Next Location Recommendation. *WWW, pp. 2177–2185*, 2021.

Dongyi Lv, Qiuyu Ding, Heng-Da Xu, Zhaoxu Sun, Zhi Wang, Feng Xiong, and Mu Xu. Reasoning Over Space: Enabling Geographic Reasoning for LLM-Based Generative Next POI Recommendation. *arXiv:2601.04562*, 2026.

Yifang Qin, Jiaxuan Xie, Zhiping Xiao, and Ming Zhang. GeoMamba: Towards Multi-granular POI Recommendation with Geographical State Space Model. *Proceedings of the AAAI Conference on Artificial Intelligence, 39(12):12479–12487*, 2025.

Yusong Wang, Chuang Yang, Jiawei Wang, Xiaohang Xu, Jiayi Xu, Dongyuan Li, Chuan Xiao, and Renhe Jiang. ELLMob: Event-Driven Human Mobility Generation with Self-Aligned LLM Framework. *ICLR*, 2026.

Wilson Wongso, Hao Xue, and Flora D. Salim. Massive-STEPS: Massive Semantic Trajectories for Understanding POI Check-ins. *arXiv:2505.11239v4; accepted to ACM SIGSPATIAL 2026*, 2026.

Xiaohang Xu, Renhe Jiang, Chuang Yang, Zipei Fan, and Kaoru Sezaki. Taming the Long Tail in Human Mobility Prediction. *Advances in Neural Information Processing Systems, 37*, 2024.

Takahiro Yabe, Kota Tsubouchi, and Toru Shimizu. Human Mobility Prediction Challenge 2024: Multi-City Prediction. *Zenodo, version 2*, 2024a.

Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, Yoshihide Sekimoto, Kaoru Sezaki, Esteban Moro, and Alex Pentland. YJMob100K: City-Scale and Longitudinal Dataset of Anonymized Human Mobility Trajectories. *Scientific Data, 11:397*, 2024b.

Dingqi Yang, Daqing Zhang, and Bingqing Qu. Participatory Cultural Mapping Based on Collective Behavior Data in Location-Based Social Networks. *ACM Transactions on Intelligent Systems and Technology, 7(3), Article 30*, 2016.

Dingqi Yang, Benjamin Fankhauser, Paolo Rosso, and Philippe Cudré-Mauroux. Location Prediction over Sparse User Mobility Traces Using RNNs: Flashback in Hidden States! *IJCAI, pp. 2184–2190*, 2020.

Song Yang, Jiamou Liu, and Kaiqi Zhao. GETNext: Trajectory Flow Map Enhanced Transformer for Next POI Recommendation. *SIGIR*, 2022.

Supplementary material. Benchmark records and figures, protocol details, and additional results are grouped in Appendices A–C.

A Benchmark Records and Supporting Figures

Mobility records use the Foursquare global resource via the AgentMove distribution (Yang et al., 2016; Feng et al., 2025a). Event sources comprise league schedules, NWS/SPC alerts, setlist.fm, archived promoter calendars, and civic records. Event records distinguish occurrence time, availability evidence, spatial binding, and provenance. An archive capture establishes a conservative availability bound at the capture time.

The multi-category inventory contains 232,005/55,113/53,540 selected training/validation/test instances (340,658 total). Closed test contains 40,221 instances, 5,641 users, and 1,315 selected episode IDs; 20,984 instances have time-of-day evidence and 19,237 have date-only evidence. The corpus has 2,555,279 training-known POIs. Model-study populations are reported separately in Appendix C.1.

Instance coverage is covered candidate instances divided by all candidate instances in the same population; event reach is reached bindings divided by admitted bindings. Covered users are distinct; these rates differ from closed-set retention.

Figures A1 and A2 summarize city–category coverage and weekly activity. Source/reporting categories describe the inventory; the broader sports, weather, and entertainment families group related records, with civic records providing supplementary coverage.

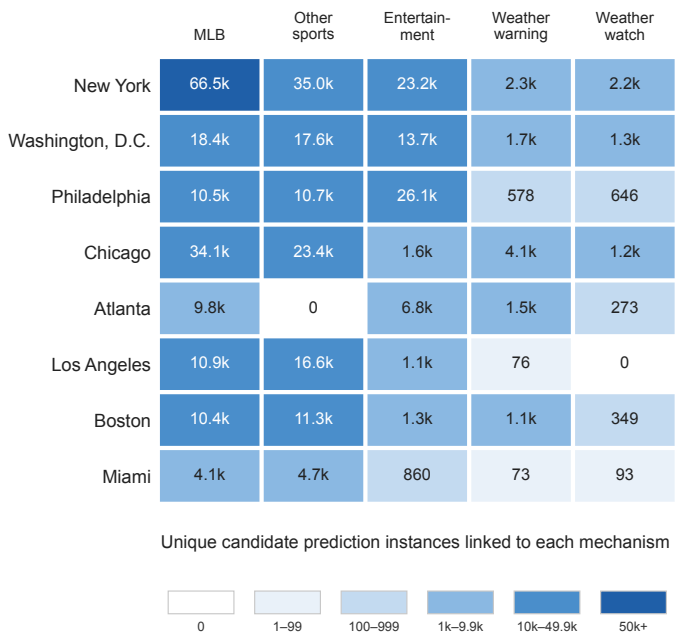

Figure A1: **City–event category coverage.** Cells count distinct linked candidate instances; category columns may overlap. Civic records are supplementary and omitted from this panel.

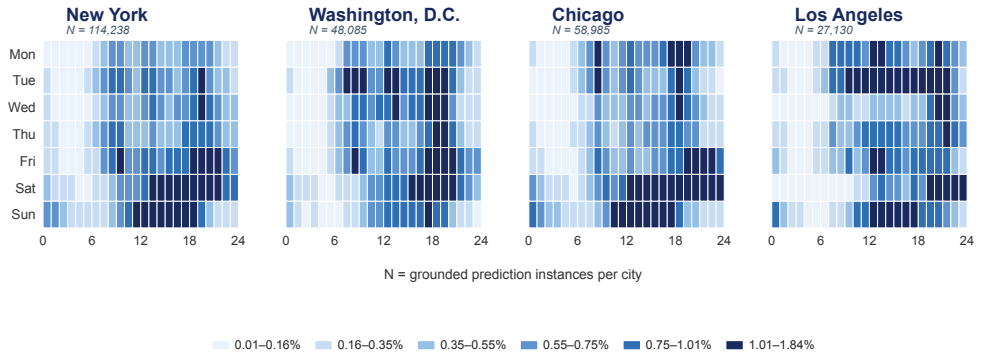

Figure A2: **Weekly activity profiles.** Each city's 168 weekday–hour cells sum to 100%; N gives its represented instance count.

B Evaluation Protocol and Model Details

B.1 Assignment, Controls, and Model Inputs

Training records define the reference city and training-known candidate vocabulary; earlier cross-city visits remain in history. The eligibility rule is Equation 2. Selected events are ordered by temporal gap and then episode identifier before target-membership filtering. Event-to-city mappings retrieve documented venue, route, or affected-area bindings, not measured physical influence boundaries. Saved assignments can be replayed by prediction-instance and episode identifiers.

T changes the supported time binding and S a family-compatible same-city place binding. Binding-derived features change with them; history, target, candidates, selected identity, and fitted model do not. Controls are not re-filtered after selection. Spatial arms share the same identity mask where needed and retain the geometry under test. The multi-category test inventory has 9,649 temporal and 10,697 spatial-control instances; the model-study sets instead have 5,761 and 11,094. The latter spatial comparison covers New York and Chicago.

Familiarity and similarity labels refer to the training event pool, with source fields and masks fixed.

EventResidual separates history and event pathways and shares candidate POI representations. Empty event input disables the correction in Equation 6; the fitted model's H pass is not an independently trained History Base. The ELLMob comparison is its next-POI adaptation, not its original trajectory-generation task.

Implementation and run configuration. The recent-history encoder consumes the latest 20 visits and uses a two-layer, four-head Transformer, while the long-history branch summarizes up to 60 active days with at most four visits per day. The event encoder uses two multi-head-attention layers with four heads and combines semantic features with explicit temporal, spatial, quality, and event-family inputs. In the v2.3 batching path, each prediction instance carries at most one selected event. Candidate-specific event residuals are bounded as $0.5\tanh(\cdot)$, and the scalar gate is disabled when the event mask is empty.

Optimization uses AdamW with learning rate 3×10^{-4}, weight decay 10^{-2}, and global gradient clipping at 1.0. The source-level objective combines event-aware sequence NLL (weight 1.0), event-masked sequence NLL (0.5), top-1/top-5/top-10 ranking margins (0.3 each), a hard-negative term (0.1), gate sparsity (0.02), event-residual L2 regularization (0.01), and general regularization (0.01). Validation uses grounded-event MRR@10 for checkpoint selection with minimum improvement 10^{-4} and early-stopping patience of three validation checks. Python, NumPy, and PyTorch RNGs are explicitly seeded, and checkpoints preserve optimizer and RNG states.

The available source confirms the settings summarized in Table A1. The implementation also exposes run-level overrides for hidden dimension, batch size, epoch limit, seed, patience, and device. The uploaded source defaults are hidden dimension 32, batch size 1, maximum 20 epochs, and seed 42, while the formal multi-seed wrapper supports seeds 42/43/44. Source-level settings are reported where run-specific overrides are unavailable; no override is inferred from defaults.

Table A1: **EventResidual implementation settings available from source.**

Item	Setting / status
Optimizer	AdamW
Learning rate / weight decay	3×10^{-4} / 10^{-2}
Gradient clipping	1.0 global norm
Recent history	Latest 20 visits
Long history	Up to 60 active days, max 4 visits/day
History encoder	2-layer, 4-head Transformer
Event encoder	2 attention layers, 4 heads
Event residual cap	0.5
Events per instance in v2.3 batching	Max 1 selected event
Validation checkpoint metric	Grounded-event (G) MRR@10
Early stopping	Min improvement 10^{-4}; patience 3 checks
RNG handling	Python / NumPy / PyTorch seeded; state checkpointed
Run-specific overrides	No override is inferred from source defaults

B.2 Metrics, Weighting, and Matching

MRR@10 averages reciprocal target ranks, with ranks above ten contributing zero; Acc@k measures whether the target is in the first k ranks. Δ_H, Δ_T, and Δ_S are paired performance effects on their named populations. For episode e and its instances Q_e, equal-episode weighting is

$$\Delta^{A,B}_{\text{episode}}(Q) = \frac{1}{|\mathcal{E}_Q|} \sum_{e \in \mathcal{E}_Q} \frac{1}{|Q_e|} \sum_{q \in Q_e} (m^A_q - m^B_q). \tag{A1}$$

On the same control set and input mask,

$$\Delta^{G,H}(Q_B) = \Delta^{G,B}(Q_B) + \Delta^{B,H}(Q_B), \qquad B \in \{T, S\}. \tag{A2}$$

Binding sensitivity is reported directly as $G - T$ and $G - S$ on their eligible paired populations; cross-population decomposition is avoided because each contrast is defined on its own evaluation set.

Absolute MRR differences are primary, with both arm scores shown. Clustered intervals are reported where directly computed, namely for the ELLMob utility populations in Appendix C.2.

Routine matching uses pre-cutoff city, calendar, and history support, without future targets or model errors. Routine references are operationally defined as instances with no selected event under the benchmark retrieval rule. Balance and reference reuse are reported in Appendix C.1.

C Additional Experimental Results

All event-conditioned results in this section use the model-study cohort of Section 5. The multi-category benchmark inventory in Appendix A is a separate reporting scope.

C.1 Population Support and Routine Matching

The model-study collection contains 291 event episodes, with 235 represented in the full test inventory of 21,966 instances and 4,735 users. Closed-set evaluation retains 18,055 instances, 4,519 users, and 233 episodes. Counts below preserve the comparison-specific populations.

Table A2: **Test populations for the event model study.** Rows overlap. The exact-time and stricter subsets are reported separately.

Population	Instances	Users	Episodes	Cities
Q_{closed}	18,055	4,519	233	8
Q_{closed}, exact time	11,734	3,056	153	6
Q_T, exact time	5,761	2,360	84	6
Q_T^{strict}	3,138	1,474	33	6
Q_S, constructible	11,094	2,260	99	2
Q_S^{hard}	4,816	1,674	41	2

Table A3: **Model-study instances by city and comparison set.** Exact denotes supported event time-of-day evidence.

City	Q_{closed}	Exact closed	Q_T	Q_T^{strict}	Q_S	Q_S^{hard}
Atlanta	1,144	1,046	585	157	0	0
Boston	977	778	502	131	0	0
Chicago	3,660	0	0	0	3,660	1,747
Los Angeles	1,269	0	0	0	0	0
Miami	220	220	213	79	0	0
New York	7,434	6,738	2,465	1,803	7,434	3,069
Philadelphia	1,049	864	628	271	0	0
Washington, D.C.	2,302	2,088	1,368	697	0	0
Total	18,055	11,734	5,761	3,138	11,094	4,816

The unseen-episode slice contains 1,384 instances, 793 users, and 172 episodes; the low-similarity slice contains 656 instances, 458 users, and 79 episodes. Both span eight cities. Table A2 gives full support; Table A4 reports equal-episode effects.

Routine-matching balance. The unmatched rate is 9.94%, with 0.0% routine-reference reuse. Standardized mean differences are 0.041 for city, 0.037 for cutoff calendar/time, 0.052 for history length, 0.028 for user support, and 0.046 for episode support. These values quantify pre-cutoff covariate balance; model-level event effects are evaluated separately through the fixed-instance comparisons in Section 5.

C.2 Paired Effects, Accuracy, and Uncertainty

Table A5 supplies accuracy for each paired population. Tables A6 and A7 report ELLMob utility with alternative aggregation and clustered intervals. The exact-time subset also has a negative $G - H$ effect.

Table A4: **EventResidual slice estimates and support.** Both weighting schemes use the named contrast on the same subset.

Subset	Contrast	Instances	Users	Episodes	Instance	Episode
Unseen episode	$G - H$	1,384	793	172	+0.0130	+0.0055
Low similarity	$G - H$	656	458	79	+0.0111	+0.0055
Strict temporal	$G - T$	3,138	1,474	33	-0.0067	-0.0065
Hard spatial	$G - S$	4,816	1,674	41	+0.0022	+0.0022

Table A5: **Supporting accuracy (%).** Each triplet is Acc@1/5/10; B names the comparison arm.

Model	Population	B	G accuracy	B accuracy
EventResidual	Q_{closed}	H	22.28 / 37.45 / 43.82	19.09 / 35.39 / 39.30
EventResidual	Q_T	T	23.69 / 38.53 / 43.41	21.10 / 36.34 / 40.41
EventResidual	Q_S	S	22.52 / 37.52 / 42.92	18.41 / 33.41 / 38.70
ELLMob (adapted)	Q_{closed}	H	17.26 / 28.52 / 33.76	18.80 / 32.10 / 36.70
ELLMob (adapted)	Q_T	T	17.76 / 30.20 / 35.46	18.71 / 29.73 / 37.16
ELLMob (adapted)	Q_S	S	16.73 / 29.45 / 37.76	17.88 / 29.88 / 36.07
ELLMob (adapted)	Exact closed	H	17.21 / 26.99 / 32.43	18.30 / 29.89 / 34.78

Table A6: **ELLMob utility under different aggregation schemes.** Values are MRR@10.

Population	Input	Instance	City macro	Episode macro
Q_{closed}	H	0.2436	0.2685	0.2715
Q_{closed}	G	0.2216	0.2348	0.2413
Exact closed	H	0.2314	0.2609	0.2671
Exact closed	G	0.2163	0.2368	0.2479

Table A7: **Reported 95% clustered intervals for ELLMob utility.** Each interval estimates $G - H$ in MRR@10.

Population	Effect	User-cluster interval	Episode-cluster interval
Q_{closed}	-0.0220	$[-0.0361, -0.0096]$	$[-0.0380, -0.0074]$
Exact closed	-0.0151	$[-0.0289, -0.0018]$	$[-0.0306, -0.0015]$

C.3 Standard Event-Off Evaluation

EventResidual-Seq is evaluated without event context on Yelp and Foursquare. GRU4Rec is a recurrent sequential recommendation comparator (Hidasi et al., 2016). These datasets and evaluation populations are separate from the event-grounding study; their absolute scores are not pooled with Table 3. Tables A8 and A9 report the sequence-prediction results.

Table A8: **Event-off MRR@10.** Best values are bold among the reported methods; NYC and TKY denote New York and Tokyo.

Dataset	MostPop	Personal freq.	Markov-1	GRU4Rec	ER-Seq
Yelp	0.0088	0.0170	0.0082	0.0057	**0.0319**
Foursquare: NYC+TKY	0.0539	0.2275	0.1536	0.1946	**0.2470**
Foursquare: NYC	0.0695	0.2570	0.1932	0.2513	**0.2749**
Foursquare: TKY	0.0476	0.2155	0.1375	0.1716	**0.2357**

Table A9: **Supporting event-off ranking metrics.** All values are fractions. PF: Personal-Frequency; ER-Seq: EventResidual-Seq.

Dataset	Model	HR@1	HR@5	HR@10	MRR@10	NDCG@10
Yelp	PF	0.0111	0.0227	0.0338	0.0170	0.0209
Yelp	ER-Seq	0.0271	0.0372	0.0594	0.0319	0.0380
Foursquare: NYC+TKY	PF	0.1502	0.3352	0.4103	0.2275	0.2710
Foursquare: NYC+TKY	ER-Seq	0.1809	0.3611	0.4317	0.2470	0.2905
Foursquare: NYC	PF	0.1628	0.3975	0.4715	0.2570	0.3083
Foursquare: NYC	ER-Seq	0.1860	0.4186	0.4926	0.2749	0.3267
Foursquare: TKY	PF	0.1451	0.3099	0.3854	0.2155	0.2559
Foursquare: TKY	ER-Seq	0.1785	0.3373	0.4069	0.2357	0.2758