TIMEARROWMOB: A CONTROLLED BENCHMARK FOR TEMPORAL DIRECTIONALITY AND IRREVERSIBILITY IN HUMAN MOBILITY AND BEYOND

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world human mobility trajectories provide rich and realistic temporal structure, but their true temporal directionality and time irreversibility are unknown, leaving existing methods without reliable ground truth for evaluation. Conversely, trajectories generated from hand-crafted rules or a single physical model can provide controlled directionality but may differ substantially from complex human mobility. We introduce TimeArrowMob, a controlled benchmark that combines real mobility structure with analytically computable temporal-direction ground truth. Taking each YJMob100K user trajectory as an independent unit, we construct a reversible baseline from its locations and symmetrized transition counts. We then impose controllable circulation constraints on transition counts and generate closed trajectories through Eulerian circuits, matching trajectory length, location frequencies, undirected transition counts, and self-transition counts across directions and strength levels. The construction generates paired Forward/Reverse trajectories with exact direction labels and additive local log-likelihood-ratio contributions relative to a fixed reference mechanism, while reversal-symmetric sampling provides zero-direction controls. The strength of the temporal arrow is explicitly controlled through the circulation scale and computed as the normalized magnitude of the summed reference contributions. Grounded in time-reversal symmetry and probabilistic inference, TimeArrowMob provides realistic trajectory morphology together with exact supervision for direction, strength, and localization. The framework enables systematic evaluation of irreversibility estimators, learning-based direction predictors, and attribution methods, and can be extended to other time-series domains with domain-valid reversal units.

1 INTRODUCTION

Next-location prediction asks a model to infer where a person will be from where they have been. Yet a correct forecast does not reveal which temporal information the model used. On trajectories dominated by long stays, repeatedly predicting the current location can yield high accuracy while ignoring the order of earlier visits. Studies of forecasting models have likewise examined how effectively they exploit temporal order (Zeng et al., 2023). We ask whether next-location prediction models understand temporal directionality in a way that shapes their predictions. We use understanding in an operational sense: the ability to exploit directional evidence in prediction and respond coherently when that evidence is changed.

Time irreversibility concerns differences between the probability distributions of trajectories and their time reversals. Reversing a recorded sequence provides an orientation label, but does not reveal how much directional evidence it contains. Seif et al. (2021) demonstrated that classifiers trained on controlled physical trajectories can recover thermodynamic evidence for time direction. Human mobility presents a different challenge: its underlying path probabilities are generally unknown. A predictor's performance gap between forward and reversed trajectories therefore cannot, by itself, establish whether it uses the relevant directional evidence.

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.

Constructing a useful benchmark requires a deliberate compromise. Preserving the original path distribution exactly would also preserve its time-reversal asymmetry, leaving no freedom to control it. CausalTime illustrates the value of combining data-derived structure with known generative ground truth for causal discovery (Cheng et al., 2024). Here, we seek controllable directional evidence while matching specified non-target statistics. Such controls address the broader concern that benchmark performance can reflect unintended cues (Geirhos et al., 2020).

We introduce TimeArrowMob, a controlled benchmark built from observed locations and undirected transition structure. Our construction builds on circulation-based perturbations of reversible Markov chains (Sun et al., 2010). Starting from a reversible baseline, we prescribe directed transition counts and realize them as closed trajectories through Eulerian circuits. Each complete trajectory receives a strict direction label and a computable strength relative to a fixed reference mechanism. Within each user, length, endpoints, location frequencies, undirected transition counts, and self-transition counts remain identical across directions and feasible strength levels. Reversal-symmetric sampling supplies zero-direction controls. Matching these statistics rules out their use as label cues, while leaving other shortcuts possible. These guarantees concern first-order directional evidence and specified statistics. Figure 1 compares the original and generated trajectories for one example user, illustrating their qualitative similarity in spatial structure.

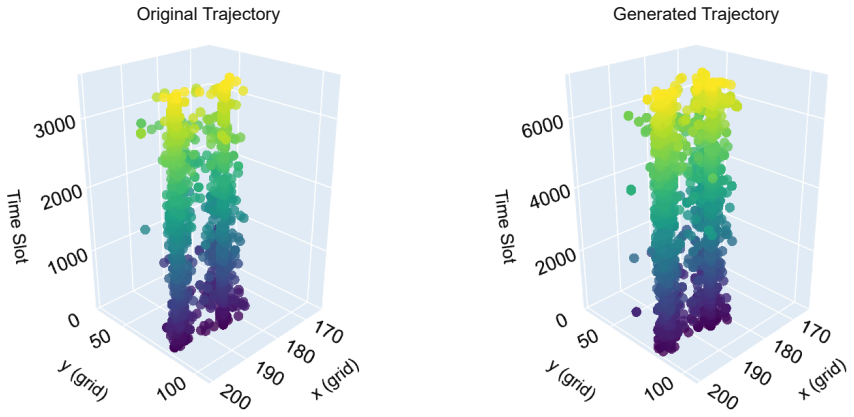

Figure 1: Original and generated trajectories for user 46. The left panel shows the preprocessed and forward-filled original trajectory (3,600 positions). The right panel shows one positive-direction trajectory at the strongest direction level (7,199 positions; directional strength: 0.05633 nat/transition). The horizontal axes show spatial grid coordinates, and each vertical axis shows its trajectory's time-slot index. Colors progress from purple to yellow as time advances. Both panels use the same spatial limits and viewing angle and display their full trajectories.

We instantiate the benchmark on YJMob100K and design a unified evaluation protocol for multiple pretrained mobility models. Parameters remain frozen, and each query produces one predicted next-location grid. Direction labels and reference weights are used only for evaluation. Forward and reversed queries are paired at the same current location. We measure prediction accuracy, directional response, and accuracy loss after reversing a closed historical segment while preserving the current location and prediction target. These measurements distinguish prediction quality, directional response, and the predictive value of historical order. We interpret them jointly as behavioral evidence of how a model uses the injected signal.

Our contributions are:

- Independent directional ground truth. A fixed reference mechanism provides exact labels and computable strengths without relying on the models being evaluated.
- Controlled trajectory generation. The construction varies first-order directional evidence while matching specified mobility statistics and providing reversible controls.
- Evaluation beyond forecast accuracy. A common protocol examines directional response and history use in frozen predictors alongside next-location accuracy.

2 Related Work

Time irreversibility has a precise statistical meaning: a stationary process is irreversible when its path distribution differs from that of its time-reversed counterpart (Lacasa et al., 2012). Stochastic thermodynamics connects forward–reverse path probabilities to entropy production under specified physical dynamics (Seifert, 2005). Practical tests examine visibility graphs (Lacasa et al., 2012; Donges et al., 2013) or ordinal patterns (Martínez et al., 2018). Their sensitivity depends on the observable, representation, sample length, and noise. Zanin & Papo (2021) demonstrate that different tests can yield complementary or conflicting assessments. Their multiscale review further examines how temporal scale and test selection affect the detection of irreversibility (Zanin & Papo, 2025). TENET learns to distinguish forward and reversed brain signals (Deco et al., 2023), while DiMViGI compares joint in- and out-degree distributions in directed multiplex visibility graphs to characterize irreversibility across multivariate interactions (Nartallo-Kaluarachchi et al., 2025). Established definitions therefore do not make irreversibility straightforward to detect in finite observations.

Mobility research already benefits from substantial real trajectory collections, including GeoLife (Zheng et al., 2011) and YJMob100K (Yabe et al., 2024). DeepMove captures sequential dependencies and periodicity through recurrent networks and historical attention (Feng et al., 2018), while Flashback revisits hidden states with relevant spatiotemporal contexts (Yang et al., 2020). GETNext combines trajectory flow graphs with Transformers (Yang et al., 2022), and NextLocLLM integrates location semantics with coordinate prediction (Liu et al., 2025). These models explicitly incorporate temporal information. Across these representative studies, however, evaluation centers on predictive accuracy and ranking; models' use of temporal direction is rarely isolated through controlled tests.

Forecasting studies also question whether predictive accuracy establishes effective use of temporal structure. Zeng et al. (2023) found that simple linear models outperformed the evaluated Transformer forecasters on long-term forecasting benchmarks and examined sensitivity to input order. Tan et al. (2024) showed that removing or replacing LLM components in three forecasting methods often preserved or improved performance, while the LLM-based methods were no more sensitive to input shuffling than their simpler ablations. These findings motivate controlled tests of temporal evidence use within the evaluated settings.

Distinguishing temporal orientations also offers a basic test of temporal representation learning. Temporal-order verification (Misra et al., 2016) and forward–backward video discrimination (Wei et al., 2018) provide self-supervised learning objectives. In physical systems, Seif et al. (2021) show that neural networks can identify the arrow of time through thermodynamically relevant observables. These studies motivate examining whether model decisions reflect directional structure, while research on shortcut learning cautions against inferring this capability from task performance alone (Geirhos et al., 2020).

TimeArrowMob connects these strands by evaluating frozen next-location predictors on controlled trajectories with known first-order directional evidence under a fixed reference mechanism. It matches specified direction-independent trajectory statistics and examines predictive accuracy alongside directional behavior, testing whether models actually use the available temporal evidence.

3 Dataset and Benchmark Recipes

Trajectory representation. For user u, a trajectory is an equally spaced location sequence $T = (s_0, \ldots, s_L)$, where $s_t \in \mathcal{V}_u$ is a discrete location identifier and L counts transitions, giving $L + 1$ observations. A fixed bijection associates each identifier with its grid coordinates. Time reversal is $T^\dagger = (s_L, \ldots, s_0)$; locations are reversed and time indices remain increasing. A closed trajectory satisfies $s_0 = s_L = o$, where o is its anchor. User and reference indices are suppressed below; all construction guarantees are conditional on a fixed user and reference mechanism.

Design objectives and scope. TimeArrowMob evaluates whether next-location predictors use temporal directionality in their predictions. We pursue five objectives: partial fidelity to observed mobility, controllable and consistent direction labels, computable trajectory-level strength, mathematical guarantees, and control of specified shortcut cues. Our main **TimeArrowMob** construction prioritizes directional control, computable ground truth, and exact label consistency, while retaining selected spatial and statistical properties of real trajectories. These priorities define an explicit com-

promise: ground truth is established independently of the evaluated predictor, without requiring preservation of the complete human mobility process.

3.1 Theoretical Foundations

Time reversal and design boundaries. For a fixed-length path distribution Q, define $Q^\dagger(T) = Q(T^\dagger)$. Reversal symmetry means $Q = Q^\dagger$; a stationary process is reversible when this equality holds at every length. Forward–reverse path probabilities provide a statistical basis for quantifying irreversibility (Roldán & Parrondo, 2012).

Irreversibility and temporal directionality. At a given path length, irreversibility detection asks whether path statistics differ under reversal ($Q \neq Q^\dagger$), establishing a distributional arrow without necessarily assigning an orientation to individual observations. In this benchmark, temporal-direction inference additionally asks whether a trajectory supports the designated forward or reverse mechanism relative to a fixed reference (Seif et al., 2021). For nonzero generated trajectories, the sign of $S_{\mathrm{ref}}(T)$ in Eq. (3.4) specifies this orientation, while $a(T)$ in Eq. (3.6) quantifies its reference strength. Irreversibility alone does not guarantee error-free orientation inference from finite or partial observations, because forward and reverse observation distributions may overlap.

Four boundaries guide our construction.

1. **Fidelity–directional-control boundary.** Preserving the complete original path distribution also fixes its reversal asymmetry. Directional control therefore requires changing some temporal joint statistics, although selected location and transition statistics can remain matched. The example in Figure 1 illustrates how spatial resemblance can be retained while temporal ordering is reconstructed.
2. **Mechanism–observation-information boundary.** Let Q_+ and $Q_- = Q_+^\dagger$ be two path distributions, and let $\mathcal{O}$ be a common observation channel independent of the direction label. Writing $\bar{Q}_\pm = \mathcal{O}Q_\pm$, data processing gives the first relation in Eq. (3.1) (Chen et al., 2016). If $\mathcal{O}$ also commutes with reversal, $\bar{Q}_- = \bar{Q}_+^\dagger$. Noise, missingness, or subsampling can consequently erase directional evidence; guarantees for complete paths do not automatically transfer to their observations.
3. **Information–direction-inference boundary.** Under equal class priors, optimal orientation accuracy satisfies the second relation in Eq. (3.1), where $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$. Known generating labels do not imply error-free inference when the observed distributions overlap.

$$D_{\mathrm{KL}}(\bar{Q}_+\|\bar{Q}_-) \le D_{\mathrm{KL}}(Q_+\|Q_-), \qquad A^* = \frac{1 + \mathrm{TV}(\bar{Q}_+, \bar{Q}_-)}{2}. \tag{3.1}$$

4. **Performance–mechanism-identification boundary.** When intended evidence and shortcut cues remain confounded, identical predictive performance can arise from different decision rules. Performance alone therefore does not identify the mechanism used by a model, motivating matched controls and history interventions (Geirhos et al., 2020).

Reversible baseline and fixed reference. Let C_{ij} count transitions $i \to j$ in the preprocessed source trajectory, including self-transitions. On its connected, nonempty support, define

$$\begin{gathered} W = C + C^\top, \qquad Z = \sum_{i,j} W_{ij}, \qquad F_0 = W/Z, \qquad \pi_i = \sum_j F_{0,ij}, \\ P_{0,ij} = F_{0,ij}/\pi_i, \qquad B = qW, \qquad L = qZ, \qquad q \in \mathbb{Z}_{>0}. \end{gathered} \tag{3.2}$$

Here F_0 is a stationary edge-flow matrix, $\pi_i > 0$ is the stationary location probability, P_0 is a transition matrix, and B is the symmetric integer count baseline. Detailed balance follows from $\pi_i P_{0,ij} = F_{0,ij} = F_{0,ji}$. Following circulation-based perturbations of reversible chains (Sun et al., 2010), choose a nonzero integer circulation J satisfying $J^\top = -J$, $\sum_j J_{ij} = 0$, and $J_{ij} = 0$

outside $E = \{(i,j) : W_{ij} > 0\}$. Such a circulation requires an undirected cycle containing at least three distinct locations. Set

$$F^{\pm} = F_0 \pm \varepsilon J, \qquad P^{\pm}_{ij} = \frac{F^{\pm}_{ij}}{\pi_i}, \qquad 0 < \varepsilon < \varepsilon_{\max} \equiv \min_{J_{ij}\neq 0} \frac{F_{0,ij}}{|J_{ij}|}. \tag{3.3}$$

The zero row and column sums of J preserve π, while $F^- = (F^+)^\top$ gives $\pi_i P^-_{ij} = \pi_j P^+_{ji}$. The reference chains therefore share π, have positive probabilities on E, and are time reversals of one another. Fix them before generating strength variants. Their stationary path probabilities are $p_\pm(T) = \pi_{s_0} \prod_{t=0}^{L-1} P^{\pm}_{s_t s_{t+1}}$. With natural logarithms and path transition counts $N_{ij}(T)$, define

$$\ell_{ij} = \log \frac{P^+_{ij}}{P^-_{ij}}, \qquad S_{\mathrm{ref}}(T) = \log \frac{p_+(T)}{p_-(T)} = \log \frac{p_+(T)}{p_+(T^\dagger)} = \sum_{(i,j)\in E} N_{ij}(T)\ell_{ij}. \tag{3.4}$$

Thus $\ell_{ji} = -\ell_{ij}$ and $S_{\mathrm{ref}}(T^\dagger) = -S_{\mathrm{ref}}(T)$.

Exact labels and strengths. For label $y \in \{-1, +1\}$, select an integer $1 \le m \le m_{\max}$ and prescribe

$$N^{(y,m)} = B + ymJ, \qquad m_{\max} = \min_{J_{ij}\neq 0} \left\lfloor \frac{B_{ij} - 1}{|J_{ij}|} \right\rfloor. \tag{3.5}$$

The zero condition uses $y = m = 0$ and $N^{(0,0)} = B$. Eq. (3.5) preserves positive counts on E, connected support, and balanced incoming and outgoing counts. An Eulerian circuit therefore realizes the prescribed counts exactly, beginning and ending at o. Since B is symmetric and ℓ is antisymmetric,

$$\kappa = \sum_{(i,j)\in E} J_{ij}\ell_{ij} > 0, \qquad S_{\mathrm{ref}}(T) = ym\kappa, \qquad a(T) = \frac{|S_{\mathrm{ref}}(T)|}{L} = \frac{m\kappa}{L}. \tag{3.6}$$

Positivity of κ follows because every nonzero J_{ij} has the same sign as ℓ_{ij}. Consequently, every generated complete path has the prescribed score sign and strength. The score and normalized strength are measured in nats and nats per transition, respectively.

For fixed reference and nonzero m, the actual output distributions Q_+ and Q_- have disjoint supports because their scores have opposite signs. Under complete noiseless observation, their Bayes orientation accuracy is one and $D_{\mathrm{KL}}(Q_+\|Q_-) = \infty$. The finite quantity $a(T)$ is therefore a **reference strength**, not the KL divergence of the actual output laws or a calibrated Bayes difficulty. It is also not claimed to recover the unknown arrow of the original human trajectory.

3.2 Task Categories and Design Rationale

Generation and matching. For each feasible m, randomized Eulerian traversal produces T^+ with counts $B+mJ$; its paired negative is $T^- = (T^+)^\dagger$. The traversal changes ordering while preserving counts and need not sample all admissible circuits uniformly. Across directions and feasible strength levels, the following invariants hold, where $V_i(T)$ counts location records:

$$V_i(T) = \sum_j B_{ij} + \mathbf{1}\{i = o\}, \qquad N_{ij}(T) + N_{ji}(T) = 2B_{ij}, \qquad N_{ii}(T) = B_{ii}. \tag{3.7}$$

Together with fixed L and o, these identities match length, endpoints, location frequencies, undirected transition counts, and self-transition counts **across generated conditions**. They do not assert exact equality with every statistic of the source trajectory. For zero-direction controls, draw a circuit U with counts B from an arbitrary law μ, then independently output U or $U^\dagger$ with equal probability:

$$Q_0(T) = \tfrac{1}{2}\mu(T) + \tfrac{1}{2}\mu(T^\dagger) = Q_0(T^\dagger). \tag{3.8}$$

This establishes reversal symmetry of the finite-path output law, beyond merely obtaining a zero reference score.

The benchmark conditions support four complementary comparisons of frozen next-location predictors (Table 1); direction labels and reference weights remain evaluation metadata.

Table 1: **Controlled comparisons in the TimeArrowMob benchmark.** Four complementary comparisons examine trajectory direction, reversal-symmetric controls, reference strength, and the predictive contribution of historical order. T^+ and T^- denote paired positive and negative trajectories, Q_0 the reversal-symmetric control distribution, and m the integer circulation scale. Direction labels and reference weights are withheld from predictors.

Comparison	Construction and purpose
Direction pairing	Pair queries from T^+ and T^- at the same current location, history length, and synthetic clock; compare predictions against their respective next locations and examine directional response.
Zero-direction control	Use Q_0 to assess prediction and history use without distributional reversal asymmetry.
Strength variation	Vary m while fixing B, J, ε, o; examine responses to different reference strengths.
Historical-order intervention	Reverse an eligible closed segment inside the visible history while preserving the current location and next-location target; assess the predictive contribution of historical order.

Complete-path labels define the comparison groups; individual history windows do not automatically inherit those labels. Section 4 defines the prediction metrics and intervention eligibility. Group mixing weights are held equal across conditions, and any data split keeps derivatives of the same source user together. Labels, filenames, and reference identifiers are excluded from predictor inputs. Directed transition counts constitute the intended first-order signal; using that signal does not establish higher-order temporal understanding.

3.3 Dataset Composition and Domain Coverage

We instantiate TimeArrowMob using Dataset 1 of **YJMob100K** (Yabe et al., 2024), which represents mobility on a 200×200 grid with 500-m cells and 30-minute time slots over 75 days. The input subset contains 1,000 source users screened for observation coverage of at least 50% and a maximum missing interval of 96 slots. The original 100 source users retain their existing preprocessing records, while the additional 900 users use the current boundary-grid cleaning rules. Forward filling produces 3,600 positions per user; imputation rates are retained as provenance. Accordingly, fidelity is assessed relative to these preprocessed inputs.

For each user, we set $q = 1$ and $\varepsilon = \varepsilon_{\max}/2$. Four distinct nonzero levels are selected as $m_r = 1 + \lfloor r(m_{\max} - 1)/3 \rfloor$, $r = 0, 1, 2, 3$, requiring $m_{\max} \geq 4$. Each feasible user contributes ten positive–negative pairs per level and ten zero-direction trajectories. Fifteen users fail this requirement under the fixed configuration, leaving 985 users and 88,650 trajectories. Every output contains 7,198 transitions and 7,199 positions. The release records seeds, reference mechanisms, realized strengths, pair relationships, and per-sample validation; empirical counts and scores are recomputed from each exported trajectory.

Coverage is currently restricted to discrete human mobility and first-order circulation evidence. Location identifiers preserve their coordinate mapping, whereas synthetic time indices do not recover original dates or circadian behavior. Cross-user strength values remain reference-dependent. Applying the recipe to other discrete sequence domains requires suitable cyclic transition support and a domain-specific fidelity assessment; broader domain validity is not established here.

4 EVALUATION

4.1 EVALUATION SETUP

Methods and evaluation scope. We evaluate seven neural predictors: NextLocLLM (Liu et al., 2025), MoVER (Lin et al., 2025), osushineko/CustomLSTM (osushinekotan, 2023), GETNext (Yang et al., 2022), ST-MoE-BERT (He et al., 2024), CityCondBERT (Oh et al., 2025), and Llama-3-8B-Mob (Tang et al., 2024)—and two rule-based methods, MOBB-style (original MOBB: Koyama et al., 2023) and CycleImputer (osushinekotan, 2023). MOBB-style denotes a local recency-weighted frequency baseline rather than the complete MOBB competition ensemble. The comparison covers nine methods without true-target input on the same 985 eligible users from the 1,000-user source cohort. Checkpoints and rule parameters remain fixed throughout evaluation. Training-user overlap is present for some checkpoints and has not been ruled out for others; these experiments assess frozen directional behavior rather than strictly held-out user generalization.

Data and queries. TimeArrowMob contains 88,650 trajectories from the 985 eligible users in the 1,000-user source cohort. Each user contributes ten zero-direction trajectories and ten positive–negative pairs at each of four nonzero strength levels. Level names indicate relative strength within users, not equal absolute strength across users. Models observe $H = 45$ consecutive half-hour positions, including the current location. We uniformly select 32 query positions per trajectory pair or zero-direction trajectory; this protocol requires 2,836,800 original predictions per method. Paired queries share the current location, history length, and synthetic clock, while each direction has its own next-location target. Predictions are single-step and are never fed back into subsequent histories. Direction labels, strengths, and reference weights are withheld from predictors. Model outputs are converted to a common grid identifier, with exact-grid accuracy as the prediction metric.

Figure 2 summarizes the paired-query evaluation and local history intervention used to measure prediction accuracy, directional response R, and historical-order benefit G.

Metrics and interventions. For one user and nonzero level, let K be the number of query pairs, i_k their shared current location, g_k^d the true next location, and $\hat{g}_k^d$ the prediction for direction $d \in \{+, -\}$. Prediction accuracy is

$$\mathrm{Acc}^d = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\{\hat{g}_k^d = g_k^d\}. \tag{4.1}$$

Using the fixed reference contributions ℓ_{ij} from Section 3, set $c = \max_{(i,j)\in E} |\ell_{ij}| > 0$ and $w_{ij} = \ell_{ij}/c$ on E, with $w_{ij} = 0$ otherwise. The same c is shared across directions and levels for each user. Directional response is

$$R = \frac{1}{2K} \sum_{k=1}^{K} \left(w_{i_k, \hat{g}_k^+} - w_{i_k, \hat{g}_k^-} \right). \tag{4.2}$$

Positive R indicates an average response aligned with the reference direction. Invalid predictions count as errors and receive zero directional contribution; self-transitions also have zero contribution.

For history intervention, we reverse the longest closed segment with nonzero reference contribution inside the visible history, breaking ties by earliest occurrence. The clock, current location, target, and location multiset remain fixed. Let I_G contain pairs eligible on both sides, $K_G = |I_G|$, and $\tilde{g}_k^d$ denote intervened predictions. The historical-order benefit is

$$G = \frac{1}{2K_G} \sum_{k \in I_G} \sum_{d \in \{+,-\}} \left[\mathbf{1}\{\hat{g}_k^d = g_k^d\} - \mathbf{1}\{\tilde{g}_k^d = g_k^d\} \right]. \tag{4.3}$$

Thus $G > 0$ means that reversal reduces accuracy on the same eligible queries. For all methods, eligibility covers 41.20%–44.44% of nonzero-level pairs and 58.43% of zero-level queries, yielding 1,257,103 additional predictions and 4,093,903 total predictions per method. Zero-direction controls

use the analogous single-sided accuracy and G; R is not applicable. Complete-trajectory labels define groups but are not assigned to individual history windows.

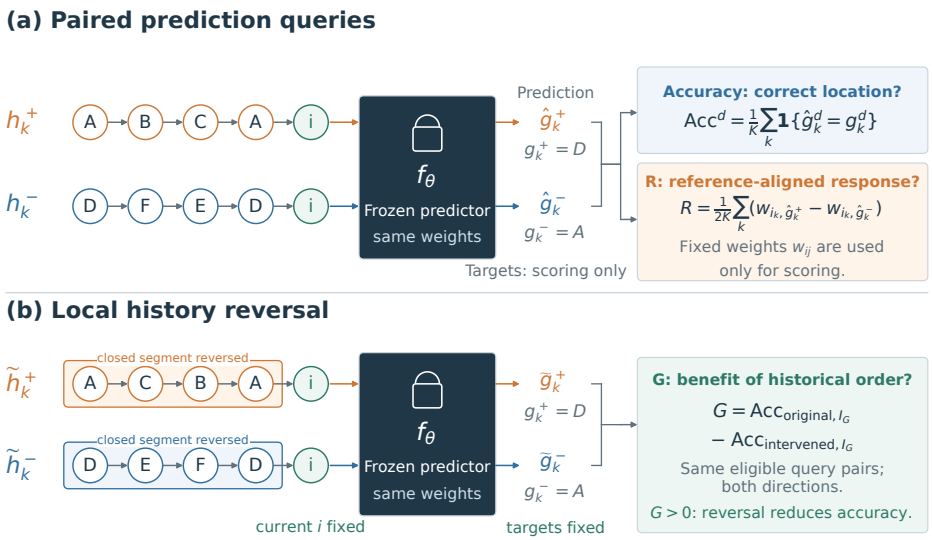

Figure 2: Paired evaluation and local history intervention for frozen predictors. (a) Paired queries from a trajectory and its complete reversal share the current location, history length, and synthetic clock, while next-location targets may differ. Original predictions determine exact-grid accuracy and directional response R. (b) Reversing the longest closed historical segment with nonzero reference contribution preserves the current location, target, clock positions, and location multiset. Historical-order benefit G is original minus intervened accuracy, pooling both directions on the same query pairs eligible on both sides. The displayed sequences are schematic; both highlighted segments are assumed to have nonzero reference contributions. Experiments use $H = 45$. Metrics are computed per user and averaged equally across users. Direction labels, strengths, reference weights, and true targets are withheld from predictors. The diagram shows the nonzero-direction protocol; zero-direction controls use eligible single queries for G.

Aggregation. We compute metrics per user and average users equally. Overall accuracy pools the nine equally sized direction–level conditions, including the zero condition. Pointwise 95% confidence intervals use 2,000 user-bootstrap replicates, without multiple-comparison correction. Accuracy is reported in percent, G in percentage points (pp), and R without units. Checkpoint details and full level-wise results appear in Appendices C and D.

4.2 Results

Prediction accuracy and directional evidence. Table 2 separates prediction quality from directional behavior. Llama-3-8B-Mob has the highest accuracy among the nine methods with completed statistics, at 40.13%, followed by ST-MoE-BERT (39.92%), MOBB-style (39.33%), MoVER (38.91%), and GETNext (35.13%). All fall below the last-location baseline (57.11%), which has $R = G = 0$. The strong persistence baseline illustrates that forecast accuracy alone cannot establish directional-information use. These accuracy differences describe the available checkpoints and adaptations, rather than a controlled comparison of architecture quality.

Small or oppositely directed responses. None of the 36 method–level R intervals for the nine completed methods has a strictly positive lower bound. MoVER exhibits small negative responses at every nonzero level. At the strongest level, its R is -0.006576 (95% CI $[-0.007248, -0.005946]$). These responses run against the reference direction. Increasing the injected strength does not yield a consistent positive response across the completed methods.

No consistent positive historical-order benefit. MoVER has positive historical-order benefits at the weak, strong, and strongest levels: $+0.307$ pp (95% CI $[+0.099, +0.515]$), $+0.444$ pp ($[+0.235, +0.649]$), and $+0.347$ pp ($[+0.083, +0.650]$), respectively; its medium-level interval in-

Table 2: **TimeArrowMob prediction accuracy and directional evidence on 985 eligible users from a 1,000-user source cohort and 88,650 trajectories.** Nine methods with completed statistics are compared with the last-location baseline. All methods exclude true-target input. Accuracy (%) is averaged equally across users and pools all nine direction–level conditions, including the zero-direction control. R denotes directional response (unitless) and G historical-order benefit (percentage points); both are averaged equally across users. Entries summarize confidence-interval categories at W / M / S / X (weak, medium, strong, and strongest), the four nonzero strength levels defined within each user. $+$ and $-$ indicate pointwise 95% confidence intervals wholly above and below zero; $\circ$ denotes any interval containing zero, including $[0, 0]$. Intervals use 2,000 user-bootstrap replicates without multiple-comparison correction. † marks CityCondBERT as a checkpoint-quality diagnostic.

Method	Accuracy (%)	R W / M / S / X	G W / M / S / X
Last-location baseline	57.11	$\circ / \circ / \circ / \circ$	$\circ / \circ / \circ / \circ$
Llama-3-8B-Mob	40.13	$\circ / - / - / -$	$\circ / \circ / \circ / \circ$
ST-MoE-BERT	39.92	$\circ / - / - / -$	$\circ / \circ / \circ / \circ$
MOBB-style (rule)	39.33	$\circ / - / - / -$	$\circ / \circ / \circ / \circ$
MoVER	38.91	$- / - / - / -$	$+ / \circ / + / +$
GETNext	35.13	$\circ / \circ / \circ / \circ$	$\circ / \circ / \circ / \circ$
osushineko/CustomLSTM	30.53	$\circ / - / - / -$	$\circ / \circ / \circ / \circ$
CycleImputer (rule)	23.39	$\circ / \circ / \circ / \circ$	$\circ / \circ / \circ / \circ$
NextLocLLM	17.44	$\circ / \circ / \circ / \circ$	$\circ / \circ / \circ / \circ$
CityCondBERT†	0.14	$\circ / \circ / \circ / \circ$	$\circ / \circ / \circ / \circ$

cludes zero. These local order benefits coexist with negative R at every nonzero level. The other eight completed methods have no G interval wholly above or below zero at any nonzero level. These unadjusted findings are exploratory; none of the completed methods demonstrates a reliably positive benefit across all four nonzero levels.

Responsiveness does not establish useful direction use. Negative R values indicate responses opposed to the reference direction and should not be described as complete insensitivity. CycleImputer has exactly zero reported G at all levels; this alone does not establish prediction invariance. CityCondBERT attains 0.1442% accuracy and has zero R at all nonzero levels, while its G intervals are either exactly zero or contain zero. Its results primarily provide a checkpoint diagnostic and are insufficient to assess the architecture's temporal capabilities.

Collectively, the nine completed methods evaluated without target information provide **no consistent evidence of jointly positive directional response and historical-order benefit across strength levels**. This conclusion concerns the present checkpoints, short histories, and single-grid prediction outputs. Intervals containing zero do not establish exact insensitivity; transfer quality, output support, and model adaptation can also affect the measurements.

5 Conclusion

We introduced TimeArrowMob, a controlled benchmark that provides exact direction labels and computable reference strengths independently of evaluated predictors, while matching selected location and transition statistics across generated conditions. Paired directions, reversal-symmetric controls, strength variation, and local history interventions distinguish prediction accuracy, directional response, and historical-order benefit.

Across 88,650 trajectories from 985 eligible users in a 1,000-user cohort, the nine methods evaluated without true-target input showed no consistent evidence of jointly positive directional response and historical-order benefit across strength levels. Several predictors responded against the reference direction; MoVER benefited from historical order at three levels despite negative directional responses. These findings motivate controlled evaluation of temporal-information use alongside forecast accuracy.

AI USE STATEMENT

AI tools were used during manuscript preparation to assist with writing and with retrieval and discovery of relevant literature.

ETHICS STATEMENT

This study uses the publicly released, anonymized YJMob100K dataset to construct synthetic trajectories for evaluating temporal directionality in next-location prediction. Although generation rearranges transitions, it retains source locations and selected mobility statistics; synthetic generation therefore does not provide a formal privacy guarantee. Potential risks include linkage to external information and inference of sensitive movement patterns. We discourage re-identification and individual surveillance, and any redistribution of data-derived artifacts should follow the source dataset's licensing and access conditions. The intended use is methodological evaluation of prediction models. The source population, user selection, and missing-data treatment may introduce biases and limit generalization. Direction labels describe a predefined reference mechanism and should not be interpreted as revealing an individual's actual behavioral intentions.

REPRODUCIBILITY STATEMENT

The code for reproducing the experiments is provided in the Supplementary Material. Section 3 specifies the mathematical definitions, construction conditions, direction labels, and reference strengths; Section 4 describes the evaluation protocol and metrics. Complete proofs, preprocessing details, and evaluation settings are provided in Appendices A–C; release records are described in Appendix E. The release will include generation parameters, random seeds, software dependencies, model and checkpoint identifiers, input adapters, query manifests, and metric-computation scripts. Validation records will enable independent checks of transition counts, statistical matching, direction scores, and reversal pairing. Original data should be obtained from its official source under the applicable terms.

REFERENCES

X. Chen, A. Guntuboyina, and Y. Zhang. On Bayes risk lower bounds. *Journal of Machine Learning Research*, 17(218):1–58, 2016. URL https://www.jmlr.org/papers/v17/16-185.html.

Y. Cheng, Z. Wang, T. Xiao, Q. Zhong, J. Suo, and K. He. CausalTime: Realistically generated time-series for benchmarking of causal discovery. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0c79d6ed1788653643a1ac67b6ea32a7-Abstract-Conference.html.

G. Deco, Y. Sanz Perl, L. de la Fuente, J. D. Sitt, B. T. T. Yeo, E. Tagliazucchi, and M. L. Kringelbach. The arrow of time of brain signals in cognition: Potential intriguing role of parts of the default mode network. *Network Neuroscience*, 7(3):966–998, 2023. doi: 10.1162/netn_a_00300.

J. F. Donges, R. V. Donner, and J. Kurths. Testing time series irreversibility using complex network methods. *EPL (Europhysics Letters)*, 102:10004, 2013. doi: 10.1209/0295-5075/102/10004.

J. Feng, Y. Li, C. Zhang, F. Sun, F. Meng, A. Guo, and D. Jin. DeepMove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*, pp. 1459–1468, 2018. doi: 10.1145/3178876.3186058.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

H. He, H. Luo, and Q. R. Wang. ST-MoE-BERT: A spatial-temporal mixture-of-experts framework for long-term cross-city mobility prediction. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Human Mobility Prediction Challenge (HuMob '24)*, pp. 10–15, 2024. doi: 10.1145/3681771.3699910.

R. Koyama, M. Suzuki, Y. Nakamura, T. Mimura, and S. Ishiguro. Estimating future human trajectories from sparse time series data. In *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge (HuMob-Challenge '23)*, pp. 26–31, 2023. doi: 10.1145/3615894.3628500.

L. Lacasa, A. Nuñez, É. Roldán, J. M. R. Parrondo, and B. Luque. Time series irreversibility: a visibility graph approach. *The European Physical Journal B*, 85:217, 2012. doi: 10.1140/epjb/e2012-20809-8.

Y. Lin, F. Lin, J. Kim, and Y.-Y. Chiang. MoVER: Modeling user heterogeneity with enriched trajectory representations for human mobility prediction. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25)*, pp. 1234–1237, 2025. doi: 10.1145/3748636.3771315.

S. Liu, N. Cao, Y. Chen, Y. Jiang, G. R. Jagadeesh, and G. Cong. NextLocLLM: Location semantics modeling and coordinate-based next location prediction with LLMs. International Workshop on Spatio-Temporal Data Intelligence and Foundation Models (STIntelligence), CIKM 2025, 2025. URL https://arxiv.org/abs/2410.09129v2. arXiv:2410.09129v2. Workshop program: https://stintelligence.github.io/.

J. H. Martínez, J. L. Herrera-Diestra, and M. Chavez. Detection of time reversibility in time series by ordinal patterns analysis. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28:123111, 2018. doi: 10.1063/1.5055855.

I. Misra, C. L. Zitnick, and M. Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *Computer Vision – ECCV 2016*, volume 9905 of *Lecture Notes in Computer Science*, pp. 527–544. Springer, 2016. doi: 10.1007/978-3-319-46448-0_32.

R. Nartallo-Kaluarachchi, L. Bonetti, G. Fernández-Rubio, P. Vuust, G. Deco, M. L. Kringelbach, R. Lambiotte, and A. Goriely. Multilevel irreversibility reveals higher-order organization of nonequilibrium interactions in human brain dynamics. *Proceedings of the National Academy of Sciences*, 122(10):e2408791122, 2025. doi: 10.1073/pnas.2408791122.

S. Oh, Y. Joo, J. Hong, O. Kwon, G. Park, D. Woo, C. Oh, H. R. Lee, and J. Heo. CityCondBERT: Cross-city transfer with GEO-BLEU-Sinkhorn loss for human mobility prediction. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25)*, pp. 1226–1229, 2025. doi: 10.1145/3748636.3771313.

osushinekotan. HuMob-Challenge-2023. GitHub software repository, 2023. URL https://github.com/osushinekotan/HuMob-Challenge-2023. CustomLSTM and CycleImputer implementations; commit d4dd05d68ee2f60df81a8d25001b16a90a79299d.

É. Roldán and J. M. R. Parrondo. Entropy production and Kullback–Leibler divergence between stationary trajectories of discrete systems. *Physical Review E*, 85(3):031129, 2012. doi: 10.1103/PhysRevE.85.031129.

A. Seif, M. Hafezi, and C. Jarzynski. Machine learning the thermodynamic arrow of time. *Nature Physics*, 17(1):105–113, 2021. doi: 10.1038/s41567-020-1018-2.

U. Seifert. Entropy production along a stochastic trajectory and an integral fluctuation theorem. *Physical Review Letters*, 95(4):040602, 2005. doi: 10.1103/PhysRevLett.95.040602.

Y. Sun, F. Gomez, and J. Schmidhuber. Improving the asymptotic performance of Markov Chain Monte-Carlo by inserting vortices. In *Advances in Neural Information Processing Systems*, volume 23, pp. 2235–2243. Curran Associates, 2010. URL https://papers.nips.cc/paper_files/paper/2010/file/819f46e52c25763a55cc642422644317-Paper.pdf.

M. Tan, M. A. Merrill, V. Gupta, T. Althoff, and T. Hartvigsen. Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems*, volume 37, pp. 60162–60191, 2024. doi: 10.52202/079017-1922.

P. Tang, C. Yang, T. Xing, X. Xu, R. Jiang, and K. Sezaki. Instruction-tuning Llama-3-8B excels in city-scale mobility prediction. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Human Mobility Prediction Challenge (HuMob '24)*, pp. 1–4, 2024. doi: 10.1145/3681771.3699908.

D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8052–8060, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wei_Learning_and_Using_CVPR_2018_paper.html.

T. Yabe, K. Tsubouchi, T. Shimizu, Y. Sekimoto, K. Sezaki, E. Moro, and A. Pentland. YJMob100K: City-scale and longitudinal dataset of anonymized human mobility trajectories. *Scientific Data*, 11:397, 2024. doi: 10.1038/s41597-024-03237-9.

D. Yang, B. Fankhauser, P. Rosso, and P. Cudré-Mauroux. Location prediction over sparse user mobility traces using RNNs: Flashback in hidden states! In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, pp. 2184–2190, 2020. doi: 10.24963/ijcai.2020/302.

S. Yang, J. Liu, and K. Zhao. 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 (SIGIR '22)*, pp. 1144–1153, 2022. doi: 10.1145/3477495.3531983.

M. Zanin and D. Papo. Algorithmic approaches for assessing irreversibility in time series: Review and comparison. *Entropy*, 23(11):1474, 2021. doi: 10.3390/e23111474.

M. Zanin and D. Papo. Algorithmic approaches for assessing multiscale irreversibility in time series: Review and comparison. *Entropy*, 27(2):126, 2025. doi: 10.3390/e27020126.

A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are Transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Y. Zheng, H. Fu, X. Xie, W.-Y. Ma, and Q. Li. *Geolife GPS trajectory dataset – User Guide*. Microsoft Research, Geolife GPS trajectories 1.1 edition, 2011. URL https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/.

A Theoretical Guarantees

A.1 Scope and assumptions

All results below are conditional on one source user and a fixed reference mechanism. The location set contains only active locations, the undirected support of $W = C + C^\top$ is connected and nonempty, and the anchor o belongs to this support. We use exactly the quantities defined in Eqs. (3.2)–(3.6): $B = qW$, $L = qZ$, an integer antisymmetric circulation J, and $0 < \varepsilon < \varepsilon_{\max}$. A nonzero strength condition is generated only when its integer scale satisfies $1 \le m \le m_{\max}$. No assumption of uniform Eulerian sampling is required.

We distinguish the stationary Markov reference laws $p_\pm$ from the actual, anchored Eulerian output laws $Q_\pm$. Reference likelihood ratios are evaluated only on paths whose transitions lie in E. Every generated complete trajectory satisfies this condition. The guarantees below concern finite discrete paths, exact arithmetic, and the specified first-order reference; numerical validation and model evaluation are separate procedures.

A.2 Precise meaning of the four design boundaries

Fidelity–directional control. Let P be the complete source path distribution at a fixed length. If a family of generated laws satisfies $Q_\theta = P$ for every control parameter θ, then $Q_\theta^\dagger = P^\dagger$ as well. Consequently, every functional of the pair $(Q_\theta, Q_\theta^\dagger)$ is fixed. In particular, total-variation asymmetry and forward–reverse KL divergence cannot vary. Directional control of the full distribution therefore requires changing that distribution. This statement does not preclude preserving selected marginals or statistics: Eqs. (3.5) and (3.7) give such a construction. It is a boundary under exact distributional fidelity, rather than a universal quantitative trade-off or an impossibility theorem for every set of design objectives.

Mechanism–observation information. Let the same direction-independent observation channel $\mathcal{O}$ act on Q_+ and $Q_- = Q_+^\dagger$. Introduce the joint laws $M_\pm(T, x) = Q_\pm(T)\mathcal{O}(x \mid T)$, whose observation marginals are $\bar{Q}_\pm$. When the joint divergence is finite, the KL chain rule gives

$$\begin{aligned} D_{\mathrm{KL}}(Q_+\|Q_-) &= D_{\mathrm{KL}}(M_+\|M_-) \\ &= D_{\mathrm{KL}}(\bar{Q}_+\|\bar{Q}_-) + \mathbb{E}_{x\sim\bar{Q}_+} D_{\mathrm{KL}}(M_+(T \mid x)\|M_-(T \mid x))\,. \end{aligned} \tag{A.1}$$

The remaining term is nonnegative, proving Eq. (3.1); if the right-hand input divergence is infinite, the inequality remains valid in the extended sense. Information loss is possible, not inevitable: an invertible observation can retain all evidence. If the channel additionally satisfies $\mathcal{O}(x^\dagger \mid T^\dagger) = \mathcal{O}(x \mid T)$, substitution of $T^\dagger$ in the marginalization proves $\bar{Q}_-(x) = \bar{Q}_+(x^\dagger)$. Direction independence alone does not imply this reversal relation.

Information–direction inference. Under equal priors, the Bayes classifier selects the larger observed probability at each observation. For discrete observations,

$$A^* = \frac{1}{2}\sum_x \max\{\bar{Q}_+(x), \bar{Q}_-(x)\} = \frac{1}{2} + \frac{1}{4}\sum_x |\bar{Q}_+(x) - \bar{Q}_-(x)| = \frac{1 + \mathrm{TV}(\bar{Q}_+, \bar{Q}_-)}{2}. \tag{A.2}$$

The same result follows with densities for general observations. Thus identical laws give chance accuracy, and disjoint supports give perfect accuracy. A known generating label does not remove overlap after an observation channel. This is an orientation-classification statement, not a formula for next-location prediction accuracy.

Performance–mechanism identification. Consider a binary intended feature X_{int} and a shortcut feature X_{shortcut} satisfying $X_{\mathrm{int}} = X_{\mathrm{shortcut}} = Y$ on the evaluation distribution. The rules $f_1(X_{\mathrm{int}}, X_{\mathrm{shortcut}}) = X_{\mathrm{int}}$ and $f_2(X_{\mathrm{int}}, X_{\mathrm{shortcut}}) = X_{\mathrm{shortcut}}$ produce identical predictions and perfect accuracy there, but respond differently when X_{shortcut} changes independently of X_{int}. Thus performance does not generally identify the feature used. Exact matching removes the specified matched statistics as direction cues conditional on the user; it does not exclude other cues or uniquely

identify an internal mechanism. The controls and interventions in Section 4 provide additional behavioral evidence under their stated conditions.

A.3 Reversibility and the reference likelihood ratio

Proposition A.1. P_0 is a reversible transition matrix with stationary distribution π, and P^+ and P^- are mutually time-reversed stationary chains.

Proof. Symmetry and nonnegativity of W imply the same properties for F_0, with total mass one. Each active row has positive mass π_i, so P_0 is row-stochastic. Moreover, $\pi_i P_{0,ij} = F_{0,ij} = F_{0,ji} = \pi_j P_{0,ji}$, proving detailed balance and stationarity. Antisymmetry and zero row sums of J imply zero column sums. The strict bound on ε ensures $F^{\pm}_{ij} > 0$ on E, and both row and column sums remain π. Hence $P^{\pm}$ are transition matrices with stationary distribution π. Finally,

$$F^-_{ij} = F^+_{ji}, \qquad P^-_{ij} = \frac{\pi_j}{\pi_i} P^+_{ji}. \tag{A.3}$$

This is precisely the stationary time-reversal relation. □

For any supported path $T = (s_0, \dots, s_L)$, including an open path, Eq. (A.3) gives

$$\begin{aligned} p_-(T) &= \pi_{s_0} \prod_{t=0}^{L-1} \frac{\pi_{s_{t+1}}}{\pi_{s_t}} P^+_{s_{t+1} s_t} \\ &= \pi_{s_L} \prod_{t=0}^{L-1} P^+_{s_{t+1} s_t} = p_+(T^\dagger). \end{aligned} \tag{A.4}$$

The stationary endpoint factor is necessary for this identity on open paths. Comparing $p_+(T)$ with $p_-(T)$ cancels their common initial factor and yields Eq. (3.4). Furthermore,

$$\ell_{ij} = \log \frac{F_{0,ij} + \varepsilon J_{ij}}{F_{0,ij} - \varepsilon J_{ij}}, \qquad \ell_{ji} = -\ell_{ij}, \qquad \ell_{ii} = 0. \tag{A.5}$$

Therefore reversing a path negates its reference score. These identities concern stationary reference probabilities even though generated circuits begin at a fixed anchor; they do not assert that Eulerian sampling follows either reference chain.

A.4 Feasible circulations, Eulerian realization, and statistical matching

Existence of a nonzero circulation. On a simple undirected cycle containing at least three distinct locations, assign $+1$ to each edge following one orientation, -1 to its reverse, and zero elsewhere. This defines a nonzero integer antisymmetric matrix with zero row sums. Conversely, suppose the off-diagonal undirected support is a forest. At a leaf, the zero-row-sum condition forces its only incident circulation value to vanish. Removing leaves inductively eliminates all edges. Self-loops cannot contribute because $J_{ii} = 0$. Thus a nonzero circulation exists exactly when the off-diagonal support contains an undirected cycle.

Existence of a circulation does not by itself ensure the requested number of feasible strength levels: the integer edge budgets must also satisfy Eq. (3.5). A source user excluded by either condition is infeasible for this construction and configuration. Such exclusion does not establish that the user's actual mobility is reversible, since the construction does not characterize its unknown full path distribution.

Proposition A.2. Every feasible $N^{(y,m)}$ can be realized by a closed Eulerian trajectory at o, and all such trajectories satisfy Eq. (3.7).

Proof. Counts are integers. For every affected edge, the definition of $m_{\max}$ gives

$$N^{(y,m)}_{ij} \geq B_{ij} - m|J_{ij}| \geq 1. \tag{A.6}$$

Unaffected supported edges retain their positive baseline counts, while unsupported edges retain zero. Let $d_i = \sum_j B_{ij}$. Zero row and column sums imply

$$\sum_j N_{ij}^{(y,m)} = d_i = \sum_j N_{ji}^{(y,m)}, \qquad \sum_{i,j} N_{ij}^{(y,m)} = \sum_{i,j} B_{ij} = L. \tag{A.7}$$

The directed multigraph has equal in-degree and out-degree at every location. Every undirected supported edge has a positive count in both orientations; connected support therefore makes the directed support strongly connected. Starting at o, following unused edges cannot stop at another location, because balance ensures an unused outgoing edge whenever an unfinished walk arrives there. The first closed walk can be enlarged by inserting closed walks through unused incident edges. Connectivity ensures that this process eventually exhausts every edge, producing an Eulerian circuit with exactly L transitions. Self-loops are traversed and counted in the same way.

Each departure from i accounts for one occurrence among $s_0, \ldots, s_{L-1}$. Including the terminal anchor gives $V_i(T) = d_i + \mathbf{1}\{i = o\}$. For distinct locations, antisymmetry cancels the circulation in the sum of the two directed counts, while the diagonal circulation is zero:

$$N_{ij} + N_{ji} = 2B_{ij} \quad (i \neq j), \qquad N_{ii} = B_{ii}. \tag{A.8}$$

These identities, the fixed anchor, and Eq. (A.7) prove all claimed matching properties across generated conditions. □

The matched visit frequencies include the repeated terminal anchor and need not equal the source trajectory's visit frequencies. Any function solely of the matched quantities is constant across a user's generated direction and strength conditions. This does not constrain unpreserved higher-order visit patterns or traversal-specific ordering cues.

A.5 Exact reference labels and the actual output laws

Proposition A.3. Every generated trajectory has $S_{\mathrm{ref}}(T) = ym\kappa$, where $\kappa > 0$ is fixed within a user and reference mechanism.

Proof. Symmetry of B and antisymmetry of ℓ give $\sum_{i,j} B_{ij}\ell_{ij} = 0$. Substituting the prescribed counts therefore gives

$$S_{\mathrm{ref}}(T) = \sum_{(i,j)\in E} (B_{ij} + ymJ_{ij})\ell_{ij} = ym\kappa. \tag{A.9}$$

By Eq. (A.5), ℓ_{ij} has the same sign as J_{ij} whenever $J_{ij} \neq 0$. Thus every nonzero summand $J_{ij}\ell_{ij}$ is positive, and at least one exists. Hence $\kappa > 0$, $\operatorname{sign} S_{\mathrm{ref}}(T) = y$ for nonzero conditions, and $a(T) = m\kappa/L$. Zero-count-baseline trajectories have zero score. □

Keeping B, J, ε fixed makes $a(T)$ linear in m; changing the reference changes its interpretation. The score is independent of the Eulerian ordering, although its local contributions are not. For a fixed positive scale, paired generation gives $Q_-(T) = Q_+(T^\dagger)$. Since Q_+ assigns probability one to score $+m\kappa$ and Q_- to score $-m\kappa$, their supports are disjoint. Consequently,

$$\mathrm{TV}(Q_+, Q_-) = 1, \qquad A^* = 1, \qquad D_{\mathrm{KL}}(Q_+ \| Q_-) = \infty. \tag{A.10}$$

The same separation holds for mixtures over positive scales with a fixed reference. It need not survive hidden-reference pooling or lossy observations. The finite reference strength is therefore neither the actual output-law divergence nor a calibration of classification difficulty. Perfect complete-path separability also implies no corresponding guarantee for a predictor receiving only a short history.

A.6 Reversal-symmetric zero-direction controls

Let μ be any probability law on anchored circuits with counts B. Reversal preserves these counts, since $B = B^\top$. An independent fair choice between $U \sim \mu$ and $U^\dagger$ induces

$$Q_0(T) = \tfrac{1}{2}\mu(T) + \tfrac{1}{2}\mu(T^\dagger) = Q_0(T^\dagger). \tag{A.11}$$

This proves both the distributional symmetry in Eq. (3.8) and, by Proposition A.3, the zero reference score of every output. The fair random reversal is important: zero score alone does not ensure that an arbitrary sampling law assigns equal probabilities to a path and its reversal. The guarantee concerns the sampling law, not exact reversal-pair frequencies in every finite realized sample. It also does not establish a stationary stochastic process reversible at all lengths. Under reversal-commuting observation channels, its observed law remains reversal-symmetric by Appendix A.2.

A.7 Local reference attribution and closed-history reversal

For a supported trajectory, define the reference contribution of a segment by

$$A_{r:s}(T) = \sum_{t=r}^{s-1} \ell_{s_t s_{t+1}}, \qquad 0 \le r < s \le L. \tag{A.12}$$

Contributions add across adjacent segments. Suppose $s_r = s_s$, and let $\widetilde{T}$ replace $(s_r, \dots, s_s)$ by its reversal. The two segment endpoints are unchanged, so any transitions immediately outside the segment are unchanged. Internal directed counts are transposed. Consequently, the location multiset, self-transition counts, and undirected transition counts are preserved, while

$$S_{\mathrm{ref}}(\widetilde{T}) - S_{\mathrm{ref}}(T) = -2A_{r:s}(T). \tag{A.13}$$

This follows by applying antisymmetry to every internal transition. A nonzero segment contribution therefore ensures that the intervention changes reference evidence. When the segment lies inside the visible history, its reversal preserves the current location and the external next-location target, as required by Eq. (4.3). The synthetic clock remains attached to its original positions.

The resulting contribution is an exact attribution of the specified first-order reference score. It is not ground truth for human behavioral causes, higher-order temporal mechanisms, or a predictor's internal attribution. Local contributions can have either sign even on a globally positive trajectory. Moreover, reversing a closed segment can change higher-order ordering alongside first-order evidence. The measured G therefore quantifies the predictive effect of this historical-order intervention; it does not isolate a unique internal mechanism or guarantee that a model's response follows the reference contribution.

B Dataset Construction and Validation

B.1 Source cohort and preprocessing

The experiment uses a source cohort of 1,000 users from Dataset 1 of YJMob100K. Source eligibility requires observation coverage of at least 50% and a maximum missing interval of 96 half-hour slots. The original 100 source users retain their existing preprocessing records; the additional 900 users use the current boundary-grid cleaning rules. These preprocessing versions must remain identifiable in the provenance records. All comparisons between generated directions and strengths are made within a user, after preprocessing.

Forward filling yields 3,600 location records for each source user. The construction therefore begins with 3,599 observed transitions in the preprocessed sequence, including self-transitions. Symmetrization gives $Z = 2 \times 3,599 = 7,198$; with $q = 1$, every generated trajectory has $L = 7,198$ transitions and $L + 1 = 7,199$ location records. The original source sequence need not be closed. Its transition graph is used to construct closed output trajectories.

Of the 1,000 source users, 985 satisfy the requirements for four distinct nonzero strength levels under the fixed configuration. The other 15 are excluded from every evaluated condition. Consequently, 1,000 denotes the source cohort and 985 denotes the common evaluation cohort. The exclusions do not vary by model, direction, or level.

Table B.1: **Dataset construction parameters and resulting sample counts.** Strength-level labels are relative to each user's fixed reference mechanism.

Item	Setting or count
Source population	YJMob100K, Dataset 1
Spatial representation	200×200 grid; 500-m cells
Source sampling interval and duration	30 minutes; 75 days
Source users / eligible users	1,000 / 985
Preprocessed positions per user	3,600
Count multiplier	$q = 1$
Reference perturbation	$\varepsilon = \varepsilon_{\max}/2$
Nonzero levels	$m_r = 1 + \lfloor r(m_{\max} - 1)/3 \rfloor$, $r = 0, 1, 2, 3$
Four-level feasibility	$m_{\max} \geq 4$, with nonzero supported circulation
Positive–negative pairs per user and level	10
Zero-direction trajectories per user	10
Trajectories per eligible user	$4 \times 10 \times 2 + 10 = 90$
Total trajectories	$985 \times 90 = 88,650$
Positions / transitions per output	7,199 / 7,198

B.2 Construction procedure

The following procedure specifies the mathematical construction. The realized circulation, anchor, traversal choices, and random seeds are part of the generation record; the guarantees do not depend on one particular cycle-selection or traversal heuristic.

1. **Build the count baseline.** For one preprocessed user trajectory, count all consecutive transitions to obtain C, including self-transitions. Compute W, Z, F_0, π, P_0, B, and L using Eq. (3.2). Retain the fixed mapping between location identifiers and grid coordinates.

2. **Fix the reference.** Choose and record an anchor o and a supported nonzero integer circulation J satisfying the conditions of Section 3.1. Compute $\varepsilon_{\max}$ and use $\varepsilon = \varepsilon_{\max}/2$ to obtain $P^{\pm}$, ℓ, and κ. These quantities remain fixed across all strength levels of that user.

3. **Check feasibility.** Compute $m_{\max}$ from Eq. (3.5). Retain the user only if the required circulation exists and four distinct nonzero levels are feasible. Select the four m_r values in Table B.1.

4. **Generate paired directions.** At each selected m, prescribe $N^{(+1,m)} = B + mJ$. Perform ten randomized Eulerian traversals, each beginning and ending at o. For each output T^+, obtain $T^- = (T^+)^\dagger$ by exact reversal. The paired output realizes $B - mJ$ without a separate approximate reconstruction.

5. **Generate zero-direction controls.** Generate ten Eulerian circuits with counts B. For each circuit, independently choose between it and its reversal with probability $1/2$, as in Eq. (3.8). This final random orientation is part of the construction of the symmetric output law.

6. **Export and validate.** Attach the user identifier, reference identifier, direction label, level, realized reference score and strength, anchor, pairing information, and generation provenance. Recompute counts and scores from each exported sequence before accepting it.

The random Eulerian traversal provides ordering variation at fixed counts; uniform sampling over all admissible circuits is not assumed. Reversal symmetry of the zero-direction sampling law also does not require a finite realized dataset to contain each sampled sequence and its exact reverse equally often. Per-trajectory zero reference scores and distributional reversal symmetry are separate guarantees.

B.3 VALIDATION IDENTITIES

For each exported trajectory, the following identities can be checked directly against its recorded user-specific reference:

$$\begin{aligned} N(T) &= B + ymJ, & s_0 &= s_L = o, \\ \sum_{i,j} N_{ij}(T) &= L, & \sum_j N_{ij}(T) &= \sum_j N_{ji}(T) = \sum_j B_{ij}, \\ S_{\mathrm{ref}}(T) &= \sum_{t=0}^{L-1} \ell_{s_t s_{t+1}} = ym\kappa, & a(T) &= \frac{|S_{\mathrm{ref}}(T)|}{L}. \end{aligned} \tag{B.1}$$

For the zero condition, $y = m = 0$. Directional outputs must additionally satisfy $T^- = (T^+)^\dagger$, $N(T^-) = N(T^+)^\top$, and opposite reference scores with equal magnitudes. Lengths, endpoints, location frequencies, undirected transition counts, and self-transition counts must agree across the user's generated conditions. Integer-count and sequence checks are exact; floating-point score comparisons require a declared numerical tolerance in the validation configuration.

The strength names weak, medium, strong, and strongest do not define common numerical cutoffs across users. Their actual values depend on J, ε, and the user's count structure. Analyses across users therefore use the recorded strengths and the within-user level definitions, rather than interpreting a shared level name as an identical physical quantity.

B.4 PRESERVED FEATURES AND INTERPRETATION

The construction retains the source location set and undirected transition support, and matches the selected count statistics across generated conditions. A generated transition uses an edge present in the symmetrized support; its orientation may be absent from the original recorded direction. Moreover, doubling and rearranging source transitions does not preserve the original ordering, dwell-time distribution, circadian pattern, or higher-order visitation motifs. The synthetic clock is a common model-input coordinate, not a reconstruction of the source dates.

These distinctions define the intended fidelity claim: the benchmark preserves specified spatial and statistical structure while providing independently established first-order directional ground truth. It does not reproduce the complete distribution of human mobility. Local contributions in Appendix A are reference-based annotations; they do not identify the causes of a user's original behavior.

C EVALUATION PROTOCOL AND MODEL ADAPTATION DETAILS

C.1 PAIRED SINGLE-STEP QUERIES

The evaluation cohort comprises 985 eligible users from 1,000 source users. Each contributes ten zero-direction trajectories and ten positive–negative pairs at each of four nonzero levels. The evaluation protocol requires all methods to receive the same histories and query locations. The history length is $H = 45$ observed positions, including the current location; each pair or zero-direction trajectory supplies 32 uniformly selected query positions. This gives $985 \times 90 \times 32 = 2{,}836{,}800$ original predictions per method. Predictions are evaluated separately and are never inserted into subsequent histories.

The following indexing makes the pairing explicit without prescribing an additional sampling rule. For $T^+ = (s_0, \ldots, s_L)$ and $T^- = (T^+)^\dagger$, a valid query at index t in T^+ pairs with index $L - t$ in T^-. Provided both histories and targets exist, the inputs and targets are

$$\begin{aligned} h_k^+ &= (s_{t-H+1}, \ldots, s_t), & g_k^+ &= s_{t+1}, \\ h_k^- &= (s_{t+H-1}, \ldots, s_t), & g_k^- &= s_{t-1}, \\ i_k &= s_t. \end{aligned} \tag{C.1}$$

Here L counts transitions, as in Section 3. Eq. (C.1) describes the correspondence between recorded positions; the actual selected indices belong to the query manifest. Paired inputs have the same synthetic clock and history length. Their targets generally differ. Synthetic clocks do not recover the original calendar dates or circadian context. Labels, strengths, reference weights, target locations, filenames, and reference identifiers are excluded from predictor inputs. Models return one next-location prediction, converted to a common grid identifier before scoring. No output probabilities are required.

C.2 Local History Reversal and Eligibility

Write a visible history as $h = (x_0, \ldots, x_{H-1})$. Using the segment contribution from Appendix A.7, a candidate intervention uses indices $r < s$ satisfying

$$x_r = x_s, \qquad A_{r:s}(h) = \sum_{t=r}^{s-1} \ell_{x_t x_{t+1}} \neq 0. \tag{C.2}$$

Among eligible segments, select the largest $s-r$, breaking ties by the earliest r. Reverse the selected positions $x_r, \ldots, x_s$ and retain the original clock positions. Equal endpoints preserve the transitions adjoining the segment and keep the current location unchanged, including when the segment reaches the history endpoint. Reversal preserves the location multiset and changes the segment's reference contribution from $A_{r:s}(h)$ to $-A_{r:s}(h)$. The next-location target is never changed or used to select the intervention.

A nonzero-level query pair enters I_G only when both histories contain an eligible segment. Eq. (4.3) compares original and intervened predictions on this identical subset. Zero-direction queries are tested independently and use a single-sided accuracy difference. A zero score for a complete trajectory does not require every historical segment to have zero score, so zero-direction controls can still support this intervention.

Eligibility is an input property, independent of model predictions. For these methods, pooled coverage is 41.20%, 41.67%, 42.89%, and 44.44% at the weak, medium, strong, and strongest levels, respectively, and 58.43% for zero-direction queries. Each method therefore requires 1,257,103 additional predictions, for 4,093,903 predictions in total. G characterizes eligible queries; it is not an effect estimate for every history. A full-trajectory label or strength is not automatically the label or strength of its 45-position history.

C.3 Metrics, Aggregation, and Baselines

Accuracy, R, and G follow Eqs. (4.1)–(4.3). A positive R indicates that predicted transitions respond, on average, in the reference-aligned direction; a negative value indicates the opposite response. Predictions outside the reference support contribute zero to R, as do self-transitions. Invalid outputs count as accuracy errors and receive zero directional contribution. Consequently, a small R can reflect output support or prediction quality as well as weak directional response.

For each user, let $\mathcal{C}$ contain the zero condition and the two directions at each of four nonzero levels, giving $|\mathcal{C}| = 9$. The overall accuracy is

$$\mathrm{Acc}_{\mathrm{overall}} = \frac{1}{985} \sum_{u=1}^{985} \frac{1}{9} \sum_{\gamma \in \mathcal{C}} \mathrm{Acc}_{u,\gamma}. \tag{C.3}$$

Level-wise R and G are also user means. In particular, G is first computed using each user's eligible queries and then averaged across users with a defined value; it is not computed by pooling all eligible queries across users. For these methods, all 985 users have a defined G at every reported level. Coverage percentages instead describe the pooled fractions of eligible queries or query pairs.

Pointwise 95% intervals use 2,000 user-bootstrap replicates, resampling users with replacement and retaining their within-user observations together. The intervals are not adjusted for multiple comparisons. Accuracy is expressed as a percentage, G as percentage points, and R without units. Neither

an interval containing zero nor an exactly zero aggregate accuracy difference proves prediction invariance. R and G are reported separately rather than combined into a score.

The last-location baseline predicts $\hat{g}_k^d = i_k$ for every query. Since $\ell_{ii} = 0$, its R is exactly zero. The intervention preserves i_k, so its original and intervened predictions coincide and $G = 0$. Its accuracy therefore measures the contribution of persistence on the same evaluation queries, without requiring training or directional information.

C.4 Frozen Predictors and Adaptation Scope

The comparison contains seven neural predictors and two rule-based methods. Existing checkpoints, preprocessing statistics, and rule parameters remain frozen. Table C.1 records the available adaptation and training-overlap information; these conditions do not establish a uniformly held-out evaluation.

Table C.1: **Documented model adaptation and training-user overlap.** Counts refer to the present evaluation cohort unless otherwise stated. Unknown overlap is not treated as evidence of independence.

Method	Adaptation and scope
NextLocLLM	The original 98 eligible users appear in the checkpoint's held-out list. All 887 additional users contributed to existing normalization statistics; 153 appear in training windows, 35 in validation windows, and 66 in conventional test windows. Evaluation permits these existing users while keeping parameters and normalization fixed.
MoVER	Uses the previously fixed location profiles and adaptation settings. Results describe these transfer conditions.
osushineko/CustomLSTM	Of the current users, 161 UIDs appear in the original training cache and 28 in the test cache. The existing global scaler uses both caches. Results do not establish unseen-user generalization.
GETNext	Uses the trained unknown-user adaptation with a shared UNK user vector for all evaluation users. The cohort includes 119 training, 16 validation, and 18 test users. Applying UNK does not remove prior training overlap.
ST-MoE-BERT	Available training-user records do not establish that all 985 users were held out.
CityCondBERT	Retains the existing local checkpoint; complete training-user independence is unverified. Its low accuracy is a checkpoint diagnostic and does not establish the architecture's performance ceiling.
MOBB-style	A local recency-weighted frequency rule using only the visible history, without neural training weights. It is not the complete MOBB competition ensemble.
CycleImputer	A rule-based method using only the visible history, without neural training weights.
Llama-3-8B-Mob	Uses the existing full checkpoint with frozen parameters. The 45 visible historical observations and a location mask for the next time slot are converted to text, and a single predicted grid is generated using a single-step JSON prefix and greedy decoding. True targets, direction labels, and strengths are excluded from the input, and no additional training is performed. Results describe this checkpoint under the single-step prediction adaptation.

Checkpoint identifiers, complete adapter settings, and software and hardware versions must be specified in the accompanying release manifest rather than inferred from model names. The present comparison evaluates frozen behavioral responses under the documented conditions; it is not a controlled ranking of architectures or a demonstration of strictly unseen-user generalization.

D Full Evaluation Results

D.1 Reporting scope and overall prediction accuracy

This appendix reports the nine methods with completed performance statistics on the same 985 eligible users selected from the 1,000-user source cohort. The cohort contains 88,650 generated

trajectories. Each method makes 2,836,800 predictions from original histories and 1,257,103 predictions after eligible history interventions, for 4,093,903 predictions per method and 36,845,127 across these nine methods.

Table D.1: **Overall exact-grid prediction accuracy.** Accuracy is computed from original histories only and averaged equally across users, pooling the nine equally sized direction–level conditions. Intervened predictions are excluded. The last-location baseline always predicts the current location. Its accuracy is 57.1113%, with a pointwise 95% user-bootstrap interval of [56.5422%, 57.6722%]. Confidence intervals for the other overall accuracies are not provided in the available results summary and are therefore not added here. † identifies CityCondBERT as a checkpoint-quality diagnostic.

Method	Accuracy (%)
Last-location baseline	57.1113
Llama-3-8B-Mob	40.1326
ST-MoE-BERT	39.9210
MOBB-style (rule)	39.3263
MoVER	38.9118
GETNext (adapted)	35.1292
osushineko/CustomLSTM	30.5272
CycleImputer (rule)	23.3916
NextLocLLM	17.4413
CityCondBERT†	0.1442

In the reported evaluation, these comparisons characterize the supplied checkpoints and adaptations; they are not a controlled ranking of model architectures or a strictly held-out-user evaluation. In particular, expansion of the source cohort changes its data-quality composition, so differences from an earlier, smaller cohort should not be interpreted as model improvements.

D.2 Intervention eligibility

Table D.2: **Shared coverage of the historical-order intervention.** For a nonzero level, coverage is the percentage of query pairs for which both directions admit an eligible intervention. For the zero condition, it is the percentage of single-sided queries admitting an intervention. This table reports the eight methods other than Llama-3-8B-Mob, all of which use the same eligible queries. The final column counts users with at least one eligible query or pair for computing G at that level. The zero condition has no paired directional response R.

Level	G coverage	Users with computable G
Zero	58.43%	985
Weak	41.20%	985
Medium	41.67%	985
Strong	42.89%	985
Strongest	44.44%	985

The local intervention is defined from the visible history and the fixed reference mechanism, independently of model predictions. The same current location and next-location target are retained for each original/intervened comparison. Accordingly, G is the original-history accuracy minus the intervened-history accuracy on the same eligible queries; it is not the difference between positive- and negative-direction accuracy.

D.3 Complete level-wise directional metrics

Table D.3: **Directional response and historical-order benefit at all five levels.** Each cell reports the equally weighted user mean followed by its pointwise 95% confidence interval from 2,000 user-bootstrap replicates. R is unitless; G is expressed in percentage points (pp). N/A denotes the absence of a positive–negative pairing in the zero condition. Level names indicate within-user circulation settings, not equal absolute reference strengths across users. Numerical precision is retained from the results summary. Intervals are not adjusted for multiple comparisons.

Method	Level	R [95% CI]	G (pp) [95% CI]
Llama-3-8B-Mob	Zero	N/A	$+0.26250$ $[-0.12379, +0.64715]$
Llama-3-8B-Mob	Weak	-0.001036768 $[-0.002216212, +0.000093130]$	$+0.15697$ $[-0.33728, +0.61509]$
Llama-3-8B-Mob	Medium	-0.001820115 $[-0.002932331, -0.000764305]$	-0.03040 $[-0.37331, +0.30111]$
Llama-3-8B-Mob	Strong	-0.003652542 $[-0.004732258, -0.002569839]$	-0.02066 $[-0.50029, +0.46857]$
Llama-3-8B-Mob	Strongest	-0.005806795 $[-0.007634358, -0.004085412]$	$+0.25020$ $[-0.12903, +0.68878]$
ST-MoE-BERT	Zero	N/A	$+0.22374$ $[+0.02891, +0.41718]$
ST-MoE-BERT	Weak	$+0.000133228$ $[-0.000235904, +0.000507607]$	$+0.00549$ $[-0.22056, +0.21795]$
ST-MoE-BERT	Medium	-0.00129248 $[-0.00167096, -0.00092831]$	-0.00243 $[-0.20865, +0.21091]$
ST-MoE-BERT	Strong	-0.00244399 $[-0.0028419, -0.00205723]$	$+0.07083$ $[-0.12573, +0.26917]$
ST-MoE-BERT	Strongest	-0.00367766 $[-0.004122, -0.00325346]$	$+0.09840$ $[-0.13199, +0.33067]$
MOBB-style (rule)	Zero	N/A	$+0.06389$ $[+0.01537, +0.11064]$
MOBB-style (rule)	Weak	$+4.36579\times10^{-5}$ $[-0.000363017, +0.000468605]$	$+0.00405$ $[-0.04915, +0.05412]$
MOBB-style (rule)	Medium	-0.00154471 $[-0.00196426, -0.00114433]$	$+0.02673$ $[-0.02219, +0.07760]$
MOBB-style (rule)	Strong	-0.00243637 $[-0.00291145, -0.00198993]$	-0.00074 $[-0.04657, +0.04641]$
MOBB-style (rule)	Strongest	-0.00369836 $[-0.00420299, -0.0032084]$	$+0.03077$ $[-0.01271, +0.07482]$
MoVER	Zero	N/A	$+0.26252$ $[+0.08731, +0.44332]$
MoVER	Weak	-0.000395673 $[-0.000813781, -8.36552\times10^{-7}]$	$+0.30701$ $[+0.09921, +0.51455]$
MoVER	Medium	-0.00274272 $[-0.00321283, -0.00228751]$	$+0.17588$ $[-0.05648, +0.40774]$
MoVER	Strong	-0.00432867 $[-0.00487165, -0.00380251]$	$+0.44374$ $[+0.23466, +0.64866]$
MoVER	Strongest	-0.00657552 $[-0.00724787, -0.00594556]$	$+0.34708$ $[+0.08284, +0.64959]$
GETNext (adapted)	Zero	N/A	-0.00500 $[-0.03061, +0.02174]$
GETNext (adapted)	Weak	$+5.96283\times10^{-6}$ $[-2.06722\times10^{-5}, +3.38127\times10^{-5}]$	-0.00567 $[-0.03121, +0.02020]$
GETNext (adapted)	Medium	$+9.71813\times10^{-6}$ $[-2.8764\times10^{-5}, +4.85326\times10^{-5}]$	-0.02219 $[-0.05096, +0.00471]$
GETNext (adapted)	Strong	-3.81138×10^{-5} $[-9.4769\times10^{-5}, +8.15728\times10^{-6}]$	-0.00151 $[-0.02131, +0.01868]$
GETNext (adapted)	Strongest	-1.52169×10^{-5} $[-4.98618\times10^{-5}, +2.20234\times10^{-5}]$	$+0.01018$ $[-0.01606, +0.03892]$

Continued on next page

Table D.3 (continued)

Method	Level	R [95% CI]	G (pp) [95% CI]
osushineko/ CustomLSTM	Zero	N/A	-0.00543 $[-0.02474, +0.01276]$
osushineko/ CustomLSTM	Weak	-8.6385×10^{-6} $[-0.000448459, +0.000422226]$	-0.00004 $[-0.01766, +0.01688]$
osushineko/ CustomLSTM	Medium	-0.000994215 $[-0.00142386, -0.000574268]$	$+0.00123$ $[-0.01513, +0.01854]$
osushineko/ CustomLSTM	Strong	-0.00150592 $[-0.00196093, -0.00107824]$	-0.00446 $[-0.02396, +0.01699]$
osushineko/ CustomLSTM	Strongest	-0.00144491 $[-0.00189757, -0.0010331]$	-0.00382 $[-0.01979, +0.01018]$
CycleImputer (rule)	Zero	N/A	$+0.00000$ $[+0.00000, +0.00000]$
CycleImputer (rule)	Weak	-0.000120732 $[-0.000620948, +0.000372462]$	$+0.00000$ $[+0.00000, +0.00000]$
CycleImputer (rule)	Medium	-0.00044753 $[-0.000914212, +4.07775\times10^{-5}]$	$+0.00000$ $[+0.00000, +0.00000]$
CycleImputer (rule)	Strong	-6.76024×10^{-5} $[-0.000526271, +0.000381938]$	$+0.00000$ $[+0.00000, +0.00000]$
CycleImputer (rule)	Strongest	-0.000436158 $[-0.000949233, +4.28482\times10^{-5}]$	$+0.00000$ $[+0.00000, +0.00000]$
NextLocLLM	Zero	N/A	-0.01741 $[-0.14753, +0.10407]$
NextLocLLM	Weak	$+7.20623\times10^{-5}$ $[-0.000196835, +0.000323903]$	$+0.03169$ $[-0.13109, +0.18467]$
NextLocLLM	Medium	-7.70679×10^{-6} $[-0.000246095, +0.000261783]$	$+0.03034$ $[-0.15929, +0.22614]$
NextLocLLM	Strong	-6.90086×10^{-5} $[-0.000317684, +0.000177056]$	$+0.01116$ $[-0.16212, +0.16693]$
NextLocLLM	Strongest	-0.000150507 $[-0.00041563, +0.000122641]$	$+0.12979$ $[-0.09585, +0.40365]$
CityCondBERT†	Zero	N/A	$+0.00000$ $[+0.00000, +0.00000]$
CityCondBERT†	Weak	$+0$ $[+0, +0]$	$+0.00000$ $[+0.00000, +0.00000]$
CityCondBERT†	Medium	$+0$ $[+0, +0]$	-0.00065 $[-0.00163, +0.00000]$
CityCondBERT†	Strong	$+0$ $[+0, +0]$	-0.00001 $[-0.00093, +0.00089]$
CityCondBERT†	Strongest	$+0$ $[+0, +0]$	$+0.00000$ $[+0.00000, +0.00000]$

For the last-location baseline, $\hat{g}_k^d = i_k$. Since $w_{ii} = 0$ and the intervention preserves the current location, this baseline has $R = 0$ at every nonzero level and $G = 0$ at all five levels. Its zero-condition R is N/A.

D.4 Interpretation of the complete results

None of the 36 nonzero-level R intervals for the nine completed methods lies wholly above zero. MoVER has negative intervals at all four nonzero levels; Llama-3-8B-Mob, ST-MoE-BERT, MOBB-style, and CustomLSTM have negative intervals at the medium, strong, and strongest lev-

els. The remaining intervals contain zero or are reported as exactly zero. Negative R indicates an average response opposed to the reference direction, not complete insensitivity to input order.

MoVER shows positive G intervals at the weak, strong, and strongest levels, while its medium-level interval contains zero. The other eight completed methods have no nonzero-level G interval wholly above or below zero. Positive historical-order benefits can therefore coexist with an oppositely directed reference response. No completed method shows jointly positive R and G across all four nonzero levels.

At the zero level, ST-MoE-BERT, MOBB-style, and MoVER have positive G intervals. This does not contradict reversal symmetry of Q_0: symmetry concerns the distribution of complete trajectories, whereas G measures a predictor's response to a targeted local modification of an eligible history. A complete trajectory with zero reference score can contain closed subsegments with nonzero contributions that cancel elsewhere. Historical-order benefit in the zero condition should therefore not be interpreted as evidence that the control distribution has a nonzero global reference direction.

Intervals containing zero do not prove the absence of temporal learning. Likewise, CycleImputer's exactly zero reported G does not by itself prove that every original and intervened prediction is identical, because accuracy differences can cancel. The results support the behavioral conclusions in Section 4 under the present checkpoints, short histories, single-grid outputs, and reference mechanism; they do not establish a universal limitation of the underlying architectures.

E Reproducibility Materials and Execution Records

E.1 Code, configuration, and artifact organization

The code for reproducing the experiments is provided in the Supplementary Material. Original YJ-Mob100K data should be obtained from its official source under the applicable terms. This appendix describes the artifacts needed to connect the mathematical construction, model inputs, predictions, and reported tables.

Table E.1: **Reproduction artifacts and their roles.** Filenames listed below are identified in the experiment summary.

Artifact	Role
Generation report and user provenance	Record source selection, preprocessing versions, exclusions, observation coverage, and imputation
Reference and trajectory records	Store C, B, J, o, ε, direction labels, realized strengths, pair identifiers, seeds, and count/score validation
Query manifests and archived model inputs	Identify original and intervened histories, current locations, targets, clocks, and eligibility
Per-query prediction records	Retain original and intervened predictions, invalid-output flags, method identifiers, and query identifiers
`metrics/model_summary.csv`	Summarize method-level results
`metrics/all_models_by_user.csv`	Supply the per-user metrics used for aggregation and uncertainty estimation
Per-model `aggregate_by_level.csv`	Report level-wise metrics and realized strength ranges
`training_overlap_preflight.json`	Document overlap with checkpoint training and preprocessing populations
`cross_model_checks.json`	Record comparisons of the common evaluation inputs
`getnext_execution_override.json`	Identify the accepted GETNext run and the excluded initial execution
Per-model frozen-state and independent-check records	Support checks of fixed weights, adapters, predictions, and metric recomputation

Exact generation and evaluation seeds, checkpoint identifiers and hashes, adapter versions, dependency versions, hardware settings, and numerical tolerances must accompany the release configuration. The available aggregate report does not enumerate these values; they are not inferred from

model names or substituted with settings from a different cohort. The construction parameters documented in Table B.1 and the evaluation parameters in Appendix C remain the settings supported by the current manuscript and experiment summary.

E.2 Execution and result verification

For methods evaluated in user shards, predictions and per-user records are combined before recomputing the full-cohort estimates and confidence intervals. Confidence intervals from separate shards are not averaged. GETNext uses one frozen instance for its accepted full-cohort inference run. Its initial sharded execution was excluded after a numerical discrepancy in a CUDA-generated graph cache was detected; those diagnostic outputs are not included in the reported prediction counts.

For MoVER, the execution record reports identical weights, buffers, location profiles, and adaptation rules across shards. A difference in an archive-path field was treated as non-computational metadata when comparing fingerprints; the prediction values were not modified. The shared location-profile archives were compared by their byte hashes.

The experiment summary reports independent recomputation of accuracy, R, G, and intervention coverage from the saved predictions. Recorded checks cover unchanged targets under intervention, the visible input range, query and input identifiers, and frozen model state. For the eight methods other than Llama-3-8B-Mob, the 985 per-user input archives have matching cross-method hashes. These checks support the use of a common evaluation input set; they do not establish that all evaluation users were absent from checkpoint training.