GHOST: MANIFOLD-GUIDED ADVERSARIAL SUBSTITUTION FOR NEXT-POI LEARNING PROTECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Check-in trajectories support mobility research, but the same temporal structure can train models to predict users' next locations. Protecting their training value poses a discrete optimization problem: substitutions must respect the semantics and geography of visited places while remaining disruptive after an adversary attempts to repair them. We introduce GHOST, a trajectory perturbation framework that combines an adversarial surrogate score with the conditional likelihood of a frozen trajectory language model. An alternating optimization procedure updates substitutions as the surrogate adapts, and stochastic candidate selection couples predictive disruption with a learned preference for real mobility patterns. A Gibbs variational interpretation exposes the rule as an adversarial tilt of a trajectory prior and makes the roles of likelihood and sampling temperature explicit. Experiments on Foursquare-NYC and Foursquare-TKY compare direct training with three paired-data restoration procedures. Under the evaluated protocols, GHOST reduces direct-training acc@1 by 60.6% on NYC and 72.1% on TKY relative to its clean-training baselines, and obtains the lowest mean accuracy after bigram-conditioned restoration among the compared methods on both datasets. Existing sensitivity results further characterize the roles of prior weighting, added entropy regularization, and leakage. These findings support manifold-guided substitution as a practical design direction for trajectory learning protection, while leaving unrestricted release security and cross-model transferability open. Code is available at https://anonymous.4open.science/r/Ghostt.

1 INTRODUCTION

Public check-in data carry two forms of value: they describe aggregate mobility and provide supervision for predicting individual destinations. Next-Point-of-Interest (next-POI) models exploit this second property through regularities in users' visits and spatial context (Luo et al., 2021; Yang et al., 2022). A publisher seeking to limit predictive reuse therefore faces a problem that anonymizing identifiers alone does not directly address: how should the training signal itself be altered before it reaches a learner?

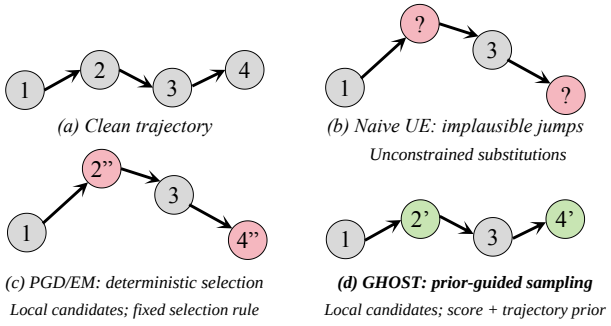

Figure 1: Conceptual comparison of substitution strategies. (a) A reference trajectory. (b) Unconstrained substitutions can produce implausible jumps. (c) PGD/EM use deterministic candidate-selection rules. (d) GHOST combines an adversarial score with a trajectory prior and samples from local candidates. The layout illustrates design choices; neither deterministic selection nor stochastic sampling alone establishes invertibility or resistance to restoration.

Unlearnable examples (UE) approach this question by modifying training data so that a model trained on the modified inputs generalizes poorly to clean inputs (Huang et al., 2021; Fu et al., 2022; Ren et al., 2023). Trajectories make this problem unusually constrained. A POI is a discrete entity with a location and a functional category, and a locally admissible substitution can still interact poorly with its temporal context. Pixel-space perturbation objectives consequently require more than a change of

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.

input representation. They require an explicit choice of feasible alternatives and a sequence-aware criterion for selecting among them. A second difficulty appears after perturbation. An adversary with clean–protected pairs can learn a restoration map and use the reconstructed data for training. BridgePure demonstrates the importance of this leakage setting for protected data (Wang et al., 2025b). In trajectories, even a simple conditional lookup table can exploit a concentrated substitution map. The relevant design question is thus not only how far accuracy falls under direct training, but how much predictive performance returns under concrete restoration procedures.

We propose **GHOST**, which couples an adversarial substitution score with the conditional likelihood supplied by a frozen trajectory language model. Candidate sets impose exact category agreement and an explicit, adaptively widened geographic neighborhood. Within each set, the adversarial score favors alternatives that weaken the surrogate's support for the genuine next destination, while the language-model score favors alternatives compatible with the current history. Sampling and surrogate adaptation connect these local choices across optimization rounds. Figure 1 summarizes this design. The central methodological choice is to couple trajectory compatibility with predictive disruption while measuring inverse recoverability separately. A likelihood model supplies a structured preference over replacements; paired-data evaluation determines whether useful training signal remains accessible to a particular restoration procedure. These are distinct questions because inversion depends on the joint clean–protected distribution, even when the outputs look natural. GHOST brings them into one experimental framework through trained learners, explicit restoration procedures, and substitution diagnostics. Our **contributions** are as follows:

- **A discrete trajectory-protection formulation.** We study next-POI training suppression under geographic and semantic candidate restrictions, distinguishing direct training from paired-data neural, unigram, and bigram restoration protocols.
- **Prior-guided adversarial substitution.** We develop a stochastic scoring rule that combines surrogate disruption and trajectory likelihood, together with alternating surrogate adaptation. Its Gibbs form exposes how score weights and temperature control candidate selection.
- **Evidence across restoration procedures.** On two category-aware mobility benchmarks, GHOST achieves a competitive combination of direct-training suppression and post-restoration accuracy. The existing ablations characterize the behavior of the prior, entropy regularization, optimization budget, and leakage without implying universal privacy guarantees.

2 RELATED WORK

Unlearnable Data. Huang et al. (2021) introduced error-minimizing perturbations that discourage a learner from acquiring useful features from protected training inputs. Subsequent work considers adversarially trained learners (Fu et al., 2022), transfer across architectures (Ren et al., 2023), and other data modalities (Liu et al., 2024; Yu et al., 2024). Jiang et al. (Jiang et al., 2024) extend UE to continuous time series. Our setting differs in the discrete output space and the geographic and semantic admissibility of substitutions. The baselines below are explicitly described as POI adaptations rather than reproductions of the original continuous-domain algorithms. Our entropy-floor variant introduces randomization at the candidate-selection stage, whereas Huang et al. (2021) and Fu et al. (2022) formulate unlearnable examples through error-minimizing perturbations.

Trajectory and POI Privacy. Never-Walk-Alone (Abul et al., 2008) studies trajectory anonymity through spatial uncertainty, while DPT (He et al., 2015) generates differentially private trajectory data. These methods address privacy definitions that differ from suppression of a downstream predictor. Preserving useful population structure is compatible with their objectives and should not itself be interpreted as a privacy failure. GHOST targets the training utility of next-POI prediction under specified protocols. It neither replaces a formal privacy mechanism nor establishes protection against reconstruction or re-identification. Geographic and category diagnostics quantify aspects of substitution quality; they do not alone establish downstream service utility.

Purification Attacks and Defences. BridgePure learns a diffusion bridge from limited clean–protected pairs to reconstruct data (Wang et al., 2025b). Its leakage perspective motivates our paired-data evaluation. Our neural restoration procedure is a discrete sequence denoiser, not an im-

plementation of the diffusion bridge. We complement it with context-free and bigram-conditioned inverse tables, which expose different forms of recoverable substitution structure. Adversarial training offers another response to delusive data (Tao et al., 2021), but is not implemented by these lookup-table procedures. The distinction matters: poor performance of one restoration model is evidence about that model and protocol, not proof that an optimal adversary has no useful inverse.

Relation to Machine Unlearning. Machine unlearning changes an already trained model to remove the influence of selected data (Liu et al., 2025). Trajectory perturbation acts before downstream training and does not require a cooperative model owner. The two approaches intervene at different points in the data lifecycle. Results on persistent poisoning influence (Pawelczyk et al., 2025; Wang et al., 2025a) further motivate studying the training input as a separate intervention point, although they do not establish the effectiveness of our method. Release-time training suppression and post-training unlearning can therefore be studied as complementary objectives, each with its own assumptions and evaluation criteria.

3 Threat Model and Preliminaries

Notation Let $\mathcal{P}$ be the POI vocabulary. A session is $s = (p_1, t_1, \ldots, p_L, t_L)$, where $p_i \in \mathcal{P}$ and t_i is its timestamp. We use one-based indexing throughout and write $s_{<i}$ for the strict prefix. A dataset contains multiple sessions per user. The publisher applies a substitution mechanism M to obtain $\tilde{s}$, with timestamps unchanged. The padding token has index zero and is excluded from substitution candidates. A next-POI example consists of a history and a target; specifying both is essential because perturbing complete sessions and perturbing only input histories yield different training protocols.

Threat Model The publisher chooses substitutions using clean training sessions and a surrogate learner. The victim trains a next-POI model, and a restoration adversary uses a leaked fraction r of paired sessions to construct alternative training data. All victim evaluations use clean held-out histories and targets. The measured outcome is predictive accuracy after training under the specified protocol; it is not a bound on all information available in a public release.

(A1) Direct protected-history training. The benchmark forms pairs $(\tilde{s}_{<i}, p_i)$: histories are substituted while supervision retains the clean next POI. This is the clean-label training-suppression setting used by the reference implementation. It measures the effect of altered histories on a prescribed learner. It does not by itself model unrestricted publication of all such labeled prefixes, whose overlap can expose clean locations.

(A2) Paired sequence denoiser. A neural denoiser learns from a fraction r of clean–protected session pairs and reconstructs non-leaked sessions. We use the existing Transformer denoiser with three argmax refinement passes. This procedure is inspired by the paired-leakage setting of BridgePure (Wang et al., 2025b); it does not implement a discrete diffusion bridge. The restored-training branch forms both histories and next-POI labels from reconstructed sessions.

(A3) Frequency-table inverter. From the same type of paired supervision, the adversary estimates

$$\hat{p}(x \mid y) = \frac{\#\{(p^{\text{clean}}, p^{\text{prot}}) = (x, y)\}}{\#\{p^{\text{prot}} = y\}}. \tag{1}$$

Each protected POI is replaced by the most frequent clean counterpart; unseen keys retain the protected POI. Training examples are then extracted from the reconstructed sessions as in A2. This baseline tests whether a context-free inverse is useful with limited paired data.

(A4) Bigram-conditioned inverter. The adversary estimates $\hat{p}(p_i^{\text{clean}} \mid p_i^{\text{prot}}, p_{i-1}^{\text{prot}})$ and falls back to the A3 table for unseen bigrams. It adds one step of protected context, providing a targeted test of sequence-dependent inversion. A4 is an adaptive extension of A3 in conditioning information, although finite-sample estimates need not improve every run.

The A1 and restored branches use different label constructions. Accordingly, differences between their accuracies describe the implemented training pipelines, not a controlled causal effect of purification alone. Comparisons *within* an attacker column use the same branch definition across methods. We retain this distinction when interpreting the aggregate metrics and discuss its implications in Section **??**.

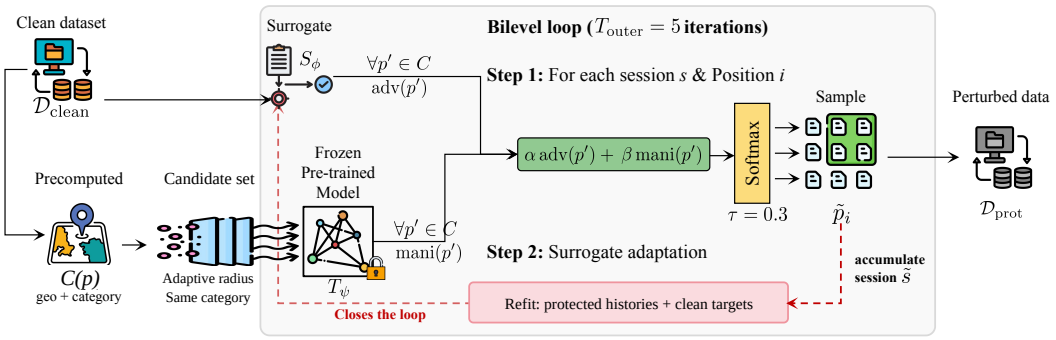

Figure 2: The GHOST framework. Geographic and category filtering constructs candidate sets with an adaptively widened radius. A surrogate supplies the adversarial score and a frozen trajectory LM supplies conditional likelihood; their weighted score drives softmax sampling. The surrogate is then refitted on protected histories with clean next-POI targets. The diagram depicts the reference implementation with five outer rounds and temperature 0.3; speed is measured diagnostically rather than enforced by a runtime mask.

Plausibility Constraints Two restrictions define the admissible alternatives. **Semantic compatibility** requires the replacement to have the original POI's category. **Geographic locality** starts from a nominal radius $R_0 = 1\,\text{km}$, with bounded widening for sparse neighborhoods. The resulting anchor-specific radius can reach $R_{\max} = 1.5^4 R_0 = 5.0625\,\text{km}$. Thus, 1 km is an initial radius rather than a universal hard displacement bound. All compared methods use this candidate-construction rule. Temporal feasibility is a separate diagnostic. We measure the fraction of consecutive released positions whose implied speed exceeds $60\,\text{km/h}$. The available implementation does not enforce a runtime speed mask, so this threshold is not claimed as a hard property of the released sequences. Category agreement, bounded candidate locality, displacement, and speed violations should be interpreted separately; satisfying one does not establish realistic mobility in every respect.

4 THE GHOST FRAMEWORK

GHOST alternates between candidate selection and surrogate adaptation. The first step scores possible substitutions using two complementary signals: how strongly the surrogate supports the clean next destination, and how likely the replacement is under a clean-trained trajectory LM. The second step refits the surrogate on the updated protected histories. Repeating the process allows the scoring model to respond to the substitutions already introduced, rather than optimizing only against a fixed clean-data predictor. The procedure is a finite alternating approximation to training-data optimization, not an exact solution of a nested optimum.

4.1 PLAUSIBILITY-FILTERED CANDIDATE SET

For each anchor p, we construct

$$C(p) = \{p' \neq p : \; d_{\text{hav}}(p, p') \leq R_p, \; \text{cat}(p') = \text{cat}(p), \; p' \neq p_{\text{pad}}\}, \tag{2}$$

retaining at most the $K = 32$ closest candidates. The radius starts at 1 km and is multiplied by 1.5 up to four times if the requested candidate count is unavailable. The reference candidate builder requests $\lceil 2^{2.5} \rceil = 6$ alternatives; this support-size setting is distinct from the optional sampling entropy floor. Category matching is never relaxed. If the maximal neighborhood remains sparse, the smaller candidate set is retained; positions with no candidates are unchanged. This construction makes the substitution support explicit, including the cost of widening sparse neighborhoods.

Candidates are shared across the compared protection methods, but their selected displacements can differ. The construction does not depend on the immediately preceding sampled point. In particular, it must not be confused with filtering an entire trajectory for a hard speed constraint.

4.2 MANIFOLD-PRIOR SUBSTITUTION SCORE

Let S_ϕ be the adapting surrogate and T_ψ a trajectory LM trained on the clean training split and then frozen. They use the same network architecture but have different roles and parameters. For position $i < L$ and candidate $p' \in C(p_i)$, define

$$\mathrm{adv}_i(p') = -z_\phi(p_{i+1}; \tilde{s}_{<i}, p'), \tag{3}$$

$$\mathrm{mani}_i(p') = \log T_\psi(p' \mid \tilde{s}_{<i}), \tag{4}$$

$$\mathrm{score}_i(p') = \alpha\, \mathrm{adv}_i(p') + \beta\, \mathrm{mani}_i(p'). \tag{5}$$

Here $z_\phi(y; h)$ is the *unnormalized output logit* for target y given history h, matching the reference implementation. It is a proxy for weakening target support, not a negative log-probability or a guarantee of increasing cross-entropy. Histories refer to the current perturbed state at scoring time, with batched updates applied during each outer round.

Adversarial term. The score favors candidates with a lower surrogate logit for the clean next POI. Its role is local and computational: it identifies substitutions that challenge the current surrogate without differentiating through a complete victim-training run. The logit scale affects its balance with the LM term, so the weights are implementation-dependent. At the final position $i = L$, the implementation uses p_{L-1} as a scoring target; this is a boundary heuristic rather than next-POI supervision. The inner training loss always uses the clean forward targets defined in A1.

Manifold term. The frozen LM supplies a conditional preference among admissible alternatives. A forward pass provides log-probabilities for all candidates at a position. For the first position, the implementation uses a one-token padding sentinel as a boundary approximation; it is not a learned unconditional first-location model. Throughout, "manifold-guided" means guidance by this likelihood, not membership in a known geometric manifold or a guaranteed high-density set.

Sampling. Substitutions are sampled according to

$$q_i(p') = \frac{\exp(\mathrm{score}_i(p')/\tau)}{\sum_{v \in C(p_i)} \exp(\mathrm{score}_i(v)/\tau)}, \qquad \tau = 0.3. \tag{6}$$

For fixed candidate scores, this distribution is the solution of

$$\max_{q \in \Delta(C)} \ \mathbb{E}_q[\alpha\, \mathrm{adv} + \beta \log T_\psi] + \tau H(q), \tag{7}$$

where H uses natural logarithms. Equation (7) gives a precise interpretation of the rule: adversarial reward and trajectory likelihood determine the preferred candidates, and temperature controls entropy regularization. There is no positive per-position entropy lower bound. A sharply concentrated score can still produce nearly deterministic substitutions.

The optional entropy-floor ablation mixes q with a uniform distribution. Its target is capped by the feasible entropy $\log_2 |C|$ (with a numerical tolerance in the implementation). The parameter measured in bits is this target entropy, not the mixing coefficient. This is an internal alternative to the default sampler rather than a property attributed to earlier UE algorithms.

Score weights. We use $(\alpha, \beta) = (2.0, 0.5)$ as the recommended configuration and $(1.0, 1.0)$ as GHOST-Sym. The asymmetric setting gives greater emphasis to surrogate disruption while retaining a trajectory-likelihood preference. Because the terms have different scales, these numbers specify an empirical operating point, not a universal ratio of protection to naturalness.

4.3 BILEVEL ALGORITHM

The inner adaptation minimizes

$$\mathcal{L}_{\mathrm{inner}}(\phi; \tilde{\mathcal{D}}, \mathcal{D}) = \sum_{s \in \mathcal{D}} \sum_{i=2}^{L_s} -\log S_\phi(p_i \mid \tilde{s}_{<i}). \tag{8}$$

The outer step uses the local scoring rule rather than the gradient of the clean-test risk through an exactly solved inner problem. We use $T_{\mathrm{outer}} = 5$ and $E_{\mathrm{inner}} = 5$, with incremental surrogate refitting. Algorithm 1 makes the clean-target dependence explicit.

Table 1: Protection and survival gaps with A1–A3 summaries. Survival gaps use within-run protected baselines; mean and worst gaps use branch-specific clean baselines (Eq. (11)).

Method	$\Delta_{\text{prot}} \uparrow$	$\Delta^{(2)}_{\text{surv}} \downarrow$	$\Delta^{(3)}_{\text{surv}} \downarrow$	$\Delta^{(4)}_{\text{surv}} \downarrow$	$\Delta_{\text{mean}} \uparrow$	$\Delta_{\text{worst}} \uparrow$
Foursquare-NYC						
GHOST	0.0835 ± 0.0032	0.0126 ± 0.0043	0.0170 ± 0.0029	0.0170 ± 0.0035	0.0743 ± 0.0019	0.0675 ± 0.0015
PGD	0.0813 ± 0.0022	0.0144 ± 0.0012	0.0188 ± 0.0052	0.0214 ± 0.0023	0.0702 ± 0.0010	0.0624 ± 0.0036
EM	0.0735 ± 0.0006	0.0102 ± 0.0044	0.0111 ± 0.0010	0.0122 ± 0.0025	0.0664 ± 0.0015	0.0615 ± 0.0012
TS-UE	0.0814 ± 0.0028	0.0093 ± 0.0045	0.0263 ± 0.0036	0.0261 ± 0.0028	0.0695 ± 0.0019	0.0551 ± 0.0053
Foursquare-TKY						
GHOST	0.1319 ± 0.0022	0.0389 ± 0.0045	0.0416 ± 0.0042	0.0519 ± 0.0065	0.1049 ± 0.0014	0.0897 ± 0.0034
PGD	0.1333 ± 0.0011	0.0397 ± 0.0013	0.0411 ± 0.0025	0.0534 ± 0.0037	0.1057 ± 0.0006	0.0911 ± 0.0020
EM	0.1116 ± 0.0016	0.0193 ± 0.0020	0.0261 ± 0.0029	0.0410 ± 0.0022	0.0964 ± 0.0004	0.0855 ± 0.0016
TS-UE	0.1135 ± 0.0005	0.0223 ± 0.0010	0.0468 ± 0.0051	0.0538 ± 0.0018	0.0898 ± 0.0014	0.0646 ± 0.0037

Matched seeds need not share protected releases; summaries are descriptive, not paired worst-case attack estimates.

For N scored positions, mean candidate count $\bar{C}$, and forward-pass cost T_{fwd}, candidate evaluation costs $O(N\bar{C}T_{\text{fwd}})$ per round. LM scoring adds $O(NT_{\text{fwd}})$, and adaptation adds $O(NE_{\text{inner}}T_{\text{fwd}})$ up to backward-pass constants. Candidate batching affects wall-clock time but not these counts. The LM is pretrained once and reused. These expressions characterize the algorithm; they are not an independently measured runtime comparison.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on Foursquare-NYC and Foursquare-TKY from TSMC2014 (Yang et al., 2015), using per-user temporal train/validation/test splits. The reference preprocessing uses an 80/10/10 allocation before rounding. POI vocabulary construction precedes splitting, making the setting transductive in POI identities; model fitting uses only training sessions. **Settings.** A causal Transformer is used as the victim, with clean-validation early stopping. The frozen trajectory LM is trained on the training split. Restoration uses 5% paired-session leakage by default, with A2 implemented as a Transformer denoising proxy. The recommended GHOST configuration is given in Section 4.2. **Baselines.** We compare against discrete POI adaptations of PGD-inspired candidate search (Madry et al., 2018), EM (Huang et al., 2021), and TS-UE (Jiang et al., 2024), sharing candidate construction and surrogate architecture. These are adaptations rather than full reproductions, and training budgets are not matched. GHOST-Sym and GHOST-EF provide internal sensitivity comparisons. Detailed preprocessing, implementation settings, and baseline definitions are provided in Appendix D. **Metrics.** We report clean-test acc@1 and protection and survival gaps relative to branch-specific clean and within-run protected baselines, respectively. Mean and worst gaps summarize A1–A3 descriptively, with A4 reported separately. Sample standard deviations are reported over three seeds where available; definit

5.2 Protection Results

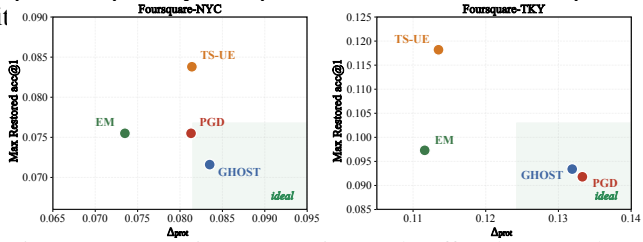

Figure 3: Protection–restoration trade-off. The axes show mean protection gap Δ_{prot} and maximum mean restored accuracy over A2–A3. The shaded lower-right region indicates the preferred direction. A4 results appear in Table 2.

Direct training is distinguished from three restoration procedures in Table 2. The main comparison is within each column: methods face the same type of learner and training-data construction. On NYC, GHOST reduces mean accuracy from 0.1379 to 0.0544; on TKY, from 0.1829 to 0.0510. These correspond to relative reductions of 60.6% and 72.1%. On NYC it achieves the lowest A1, A3, and A4 means, while TS-UE is slightly lower under A2. On TKY it achieves the lowest A2 and A4 means, with PGD lower under A1 and A3. The strength of the result is therefore the combined profile across procedures, not uniform superiority in every cell.

Table 2: Clean-test acc@1 (mean ± sample std., three seeds). Clean uses the A1 baseline; restoration results use their respective runs. Bold marks the lowest mean per dataset and outcome column.

Method	Clean (A1)	A1 ↓	A2 ↓	A3 ↓	A4 ↓
Foursquare-NYC					
GHOST	0.1379 ± 0.0009	**0.0544** ± 0.0031	0.0672 ± 0.0038	**0.0716** ± 0.0007	**0.0733** ± 0.0005
PGD	0.1379 ± 0.0009	0.0566 ± 0.0023	0.0711 ± 0.0014	0.0755 ± 0.0030	0.0771 ± 0.0020
EM	0.1379 ± 0.0009	0.0644 ± 0.0014	0.0746 ± 0.0043	0.0755 ± 0.0006	0.0775 ± 0.0013
TS-UE	0.1389 ± 0.0021	0.0575 ± 0.0020	**0.0668** ± 0.0043	0.0838 ± 0.0032	0.0836 ± 0.0020
Foursquare-TKY					
GHOST	0.1829 ± 0.0008	0.0510 ± 0.0028	**0.0899** ± 0.0017	0.0934 ± 0.0027	**0.1014** ± 0.0042
PGD	0.1836 ± 0.0009	**0.0503** ± 0.0012	0.0904 ± 0.0000	**0.0918** ± 0.0015	0.1041 ± 0.0030
EM	0.1828 ± 0.0015	0.0712 ± 0.0028	0.0905 ± 0.0011	0.0973 ± 0.0002	0.1112 ± 0.0037
TS-UE	0.1837 ± 0.0009	0.0702 ± 0.0015	0.0924 ± 0.0015	0.1182 ± 0.0049	0.1240 ± 0.0026

A2: paired sequence denoiser; A3: unigram inversion; A4: bigram inversion. A printed zero standard deviation reflects rounding.

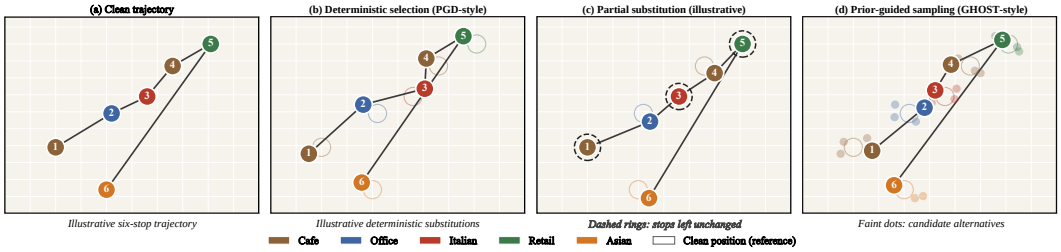

Figure 4: Conceptual six-stop trajectories contrasting deterministic selection, partial substitution, and prior-guided sampling. Colors denote POI categories; hollow rings, dashed rings, and faint dots indicate reference positions, unchanged stops, and candidate alternatives. The illustration is schematic, not an empirical evaluation of utility or inversion resistance.

Headline metric. The aggregate summaries in Table 1 complement the individual accuracy columns. Mean performance reflects an equal-weight summary over the specified A1–A3 pipelines; the worst summary takes the minimum of their own-clean gaps. The worst summary cannot exceed direct-training protection: because A1 is included, $\Delta_{\text{worst}} \leq \Delta_{\text{prot}}$. On NYC, GHOST obtains a mean own-clean gap of 0.0743 versus 0.0702 for PGD. On TKY, the corresponding means are 0.1049 and 0.1057. We retain the individual restored accuracies as the more direct comparison. With three seeds, small differences should be interpreted at the scale of observed run variability; overlapping standard deviations are not a test of equivalence. **Pareto-optimal trade-off.** Direct-training protection is plotted against maximum restored accuracy over A2–A3 in Figure 3. It makes the trade-off explicit without averaging away the attack identities. A point is Pareto-optimal only relative to these plotted means and this attack set. Nearness to another point or overlap in run variability does not place both on a strict Pareto frontier. Including A4 can change the frontier, which is why its results remain visible in the main accuracy table. **Adaptive sequence-aware adversary.** A4 conditions the inverse table on the previous protected POI as well as the current one. This tests a specific way in which a context-free attack can miss recoverable sequence structure. GHOST has the lowest mean A4-restored accuracy among the compared methods: 0.0733 on NYC and 0.1014 on TKY, compared with PGD's 0.0771 and 0.1041. The absolute differences are approximately 0.38 and 0.27 percentage points. This is an empirical ranking under limited paired supervision; it does not imply that additional context is intrinsically uninformative or that a stronger sequence model would obtain the same ranking.

5.3 Per-Method Failure Modes

Different sensitivities to direct training and restoration are observed across methods in Table 2. These outcomes characterize the evaluated pipelines rather than isolate causal mechanisms. A detailed analysis is provided in Appendix C.

Table 3: Substitution diagnostics from the A1 source records. Geo. exc. is the fraction of substitutions beyond the nominal 1 km radius; speed viol. measures adjacent released transitions above 60 km/h. Displacements are in km. All selected records have category-match rate 1.000.

Method	Sub. rate	Geo. exc.	Speed viol.	Mean disp.	P95 disp.
Foursquare-NYC					
GHOST	0.861 ± 0.000	0.645 ± 0.005	0.119 ± 0.000	1.932 ± 0.004	4.717 ± 0.009
PGD	0.861 ± 0.000	0.646 ± 0.000	0.119 ± 0.000	1.938 ± 0.013	4.713 ± 0.021
EM	0.861 ± 0.000	0.615 ± 0.004	0.118 ± 0.000	1.855 ± 0.017	4.695 ± 0.019
TS-UE	0.749 ± 0.000	0.632 ± 0.006	0.122 ± 0.001	1.909 ± 0.021	4.702 ± 0.017
Foursquare-TKY					
GHOST	0.969 ± 0.000	0.443 ± 0.002	0.070 ± 0.001	1.349 ± 0.007	4.294 ± 0.021
PGD	0.969 ± 0.000	0.443 ± 0.002	0.072 ± 0.000	1.355 ± 0.002	4.307 ± 0.017
EM	0.969 ± 0.000	0.394 ± 0.002	0.063 ± 0.001	1.187 ± 0.008	4.101 ± 0.019
TS-UE	0.855 ± 0.000	0.415 ± 0.010	0.070 ± 0.001	1.245 ± 0.024	4.169 ± 0.051

5.4 Plausibility Comparison

Per-position plausibility. Table 3 reports substitution rate, exact category agreement, displacement, and speed diagnostics for the same A1 source records used in the main comparison. The nominal-radius exceedance rate counts replacements beyond 1 km after allowed widening; it is not a violation of the larger anchor-specific search radius. The observed exceedances show that widening materially affects the operating regime. Likewise, the speed-violation rate is nonzero, so geographic/category filtering alone does not ensure feasible motion. These diagnostics characterize the released sequences more accurately than a single "plausible" label.

Sequence-level naturalness. The existing mean conditional log-likelihood measurements under the frozen LM are reported in Table A1. GHOST improves the mean score relative to PGD from -10.3664 to -10.0208 on NYC and from -9.7827 to -9.2804 on TKY. TS-UE and EM attain higher likelihood than GHOST on TKY, showing that this diagnostic and predictive disruption rank alternatives differently. The clean sequences provide a reference, not a mathematical upper bound, and the substantial gap to that reference remains. Because the same LM guides optimization and measures likelihood, this is an internal alignment diagnostic rather than an independent validation of realistic behavior or preserved service utility.

5.5 Ablation Study

Table 4: Existing NYC variant comparison (three seeds). GHOST-Sym uses symmetric score weights. GHOST-EF is the internal non-LM variant with a saved 2.5-bit target and temperature 0.7, so it is not an isolated ablation of the prior.

Variant	A1 ↓	A2 ↓	A3 ↓
GHOST	0.0544 ± 0.0031	0.0672 ± 0.0038	0.0716 ± 0.0007
GHOST-Sym	0.0619 ± 0.0027	0.0693 ± 0.0018	0.0715 ± 0.0009
GHOST-EF	0.0666 ± 0.0004	0.0762 ± 0.0026	0.0712 ± 0.0036

Manifold-prior variants. The recommended weights, symmetric weights, and internal entropy-floor variant are compared in Table 4. The recommended configuration has stronger direct-training suppression than the symmetric variant in the existing results. Rankings need not agree across restoration procedures. Since the variants also differ in score balance or sampling regularization, this comparison supports the combined operating point rather than attributing the entire gain to the LM term alone.

Score weighting (α, β) The existing weight sweep in Table A2 varies both adversarial and likelihood contributions. It illustrates a practical trade-off: increasing the likelihood weight can favor easier-to-predict substitutes, while stronger adversarial weighting can reduce direct-training accuracy. Changing (α, β) also changes the scale of the score relative to temperature, so the sweep should not be interpreted as a pure experiment on one isolated mechanism. The recommended configuration is a useful observed operating point, not a proven optimum.

Entropy floor Table A3 summarizes the available entropy-target sweep with the recommended score weights. An explicit entropy target changes which candidates are sampled and can dilute

high-scoring substitutions. The results support retaining the default without an added floor in this sweep; they do not establish that the LM itself enforces high entropy. The response is not assumed to be monotonic, and a bit-valued target is capped by the entropy capacity of each candidate set.

Bilevel convergence Figure A1 reports sensitivity to the number of outer rounds. The mean summary varies within a relatively narrow band in the existing runs, whereas the worst summary varies more. We describe this as empirical budget sensitivity rather than a convergence proof: similar downstream metrics do not show that model parameters or substitution distributions have converged. Five rounds remain the chosen default in the evaluated configuration.

Leak ratio sensitivity Figure 5 summarizes the existing leakage sweep from 1% to 20%. More paired information can improve restoration, but finite-sample training and changes in the protected realization can make the curve nonmonotonic. At very low leakage, negative survival gaps indicate that the implemented reconstruction-and-training pipeline performs worse than its paired protected-history run. An adversary allowed to select its best pipeline need not accept this deterioration. The relative ranking also depends on both leakage and the restoration procedure, so the sweep does not establish uniform resistance for all leakage levels.

5.6 Computational Cost

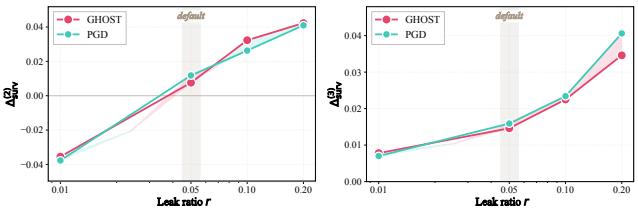

Figure 5: Leakage sensitivity on NYC. Survival gaps measure recovery relative to each run's protected baseline, not absolute restored accuracy. A2 is a Transformer denoising proxy; A3 is unigram inversion. Lower is better; shading shows curve separation, not uncertainty.

The extra scoring component in GHOST is one LM forward pass per history, from which all candidate likelihoods are gathered. Relative to evaluating the surrogate on every candidate, this adds an $O(1/\bar{C})$ forward-pass component before accounting for batching, surrogate refitting, and LM pretraining. The practical cost therefore depends on candidate cardinality and implementation throughput. Reusing the frozen LM amortizes its training across configurations. The saved results do not provide a controlled, hardware-matched runtime study, so we do not infer measured wall-clock superiority or equal compute budgets from asymptotic counts.

5.7 Understanding Manifold-Guided Purification Resistance

For a fixed history, let π_C denote the LM distribution renormalized over the candidate set. The sampling rule becomes

$$q(p') \propto \exp\left(\frac{\alpha\,\mathrm{adv}(p')}{\tau}\right)\pi_C(p')^{\beta/\tau}. \tag{9}$$

Thus, GHOST applies an adversarial tilt to a temperature-adjusted trajectory prior, combining surrogate disruption with likelihood-guided selection. This interpretation does not establish irreversibility: recovery depends on the joint clean–protected distribution. The reported resistance is empirical and specific to the evaluated restoration procedures. Further analysis is provided in Appendix F. Deployment considerations, and evaluation limitations are discussed in Appendix G.

6 Conclusion

GHOST formulates trajectory perturbation as prior-guided adversarial substitution over discrete POI candidates. Its alternating optimization and stochastic sampling combine surrogate disruption with a learned preference for mobility structure. Across the evaluated Foursquare benchmarks, the results show competitive direct-training suppression and post-restoration performance, including favorable mean accuracy under bigram-conditioned inversion. These findings motivate evaluating protection jointly with substitution quality and explicit restoration procedures, while distinguishing likelihood alignment from irreversibility. GHOST provides a concrete framework and empirical foundation for trajectory learning protection; security under unrestricted release remains an open question.

AI USE STATEMENT

AI tools assisted with language checking and polishing. No additional model-training or inference experiments were performed as part of these revisions. The authors take responsibility for the accuracy of the reported results, citations, and final manuscript.

ETHICS STATEMENT

Check-in data can reveal sensitive behavior. This work examines suppression of next-POI learning under specified benchmark protocols, and the schematic figures do not expose individual user trajectories. Lower prediction accuracy is not a formal privacy guarantee and does not establish resistance to reconstruction, re-identification, or inference through released labels. Category agreement and local substitution do not by themselves establish downstream utility. Any deployment should be evaluated for its actual release interface and affected users rather than inferred from the benchmark alone.

REPRODUCIBILITY STATEMENT

Code is available at https://anonymous.4open.science/r/Ghostt. The repository includes a reference training configuration, saved prediction records, and scripts for recomputing statistical summaries. Experimental settings and evaluation protocols are documented in the main text and appendix.

REFERENCES

Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *Proceedings of the 24th IEEE International Conference on Data Engineering*, pp. 376–385. IEEE, 2008. doi: 10.1109/ICDE.2008.4497446.

Shaopeng Fu, Fengxiang He, Yang Liu, Li Shen, and Dacheng Tao. Robust unlearnable examples: Protecting data against adversarial learning. In *10th International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=baUQQPwQiAg.

Xi He, Graham Cormode, Ashwin Machanavajjhala, Cecilia M. Procopiuc, and Divesh Srivastava. DPT: Differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, 8(11):1154–1165, 2015. doi: 10.14778/2809974.2809978.

Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *9th International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=iAmZUo0DxC0.

Yujing Jiang, Xingjun Ma, Sarah Monazam Erfani, and James Bailey. Unlearnable examples for time series. In *Advances in Knowledge Discovery and Data Mining*, volume 14650 of *Lecture Notes in Computer Science*, pp. 213–225. Springer, 2024. doi: 10.1007/978-981-97-2266-2_17.

Xinwei Liu, Xiaojun Jia, Yuan Xun, Siyuan Liang, and Xiaochun Cao. Multimodal unlearnable examples: Protecting data against multimodal contrastive learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8024–8033. ACM, 2024. doi: 10.1145/3664647.3680708.

Ziyao Liu, Huanyi Ye, Chen Chen, Yongsen Zheng, and Kwok-Yan Lam. Threats, attacks, and defenses in machine unlearning: A survey. *IEEE Open Journal of the Computer Society*, 6:413–425, 2025. doi: 10.1109/OJCS.2025.3543483.

Yingtao Luo, Qiang Liu, and Zhaocheng Liu. STAN: Spatio-temporal attention network for next location recommendation. In *Proceedings of the Web Conference 2021*, pp. 2177–2185. ACM, 2021. doi: 10.1145/3442381.3449998.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Martin Pawelczyk, Jimmy Z. Di, Yiwei Lu, Gautam Kamath, Ayush Sekhari, and Seth Neel. Machine unlearning fails to remove data poisoning attacks. In *13th International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7e810b2c75d69be186cadd2fe3febeab-Abstract-Conference.html.

Jie Ren, Han Xu, Yuxuan Wan, Xingjun Ma, Lichao Sun, and Jiliang Tang. Transferable unlearnable examples. In *11th International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=-htnolWDLvP.

Lue Tao, Lei Feng, Jinfeng Yi, Sheng-Jun Huang, and Songcan Chen. Better safe than sorry: Preventing delusive adversaries with adversarial training. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16209–16225, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/8726bb30dc7ce15023daa8ff8402bcfd-Abstract.html.

Wenbin Wang, Qiwen Ma, Zifan Zhang, Yuchen Liu, Zhuqing Liu, and Minghong Fang. Poisoning attacks and defenses to federated unlearning. In *Companion Proceedings of the ACM Web Conference 2025*, pp. 1365–1369. ACM, 2025a. doi: 10.1145/3701716.3715494.

Yihan Wang, Yiwei Lu, Xiao-Shan Gao, Gautam Kamath, and Yaoliang Yu. BridgePure: Limited protection leakage can break black-box data protection. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-0950. URL https://papers.neurips.cc/paper_files/paper/2025/hash/28f2462d39136cac2a2e7ff313e444a7-Abstract-Conference.html.

Dingqi Yang, Daqing Zhang, Vincent W. Zheng, and Zhiyong Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2015. doi: 10.1109/TSMC.2014.2327053.

Song Yang, Jiamou Liu, and Kaiqi 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*, pp. 1144–1153. ACM, 2022. doi: 10.1145/3477495.3531983.

Yi Yu, Qichen Zheng, Siyuan Yang, Wenhan Yang, Jun Liu, Shijian Lu, Yap-Peng Tan, Kwok-Yan Lam, and Alex Kot. Unlearnable examples detection via iterative filtering. In *Artificial Neural Networks and Machine Learning – ICANN 2024*, volume 15025 of *Lecture Notes in Computer Science*, pp. 241–256. Springer, 2024. doi: 10.1007/978-3-031-72359-9_18.

A Appendix

B Additional Figures and Tables

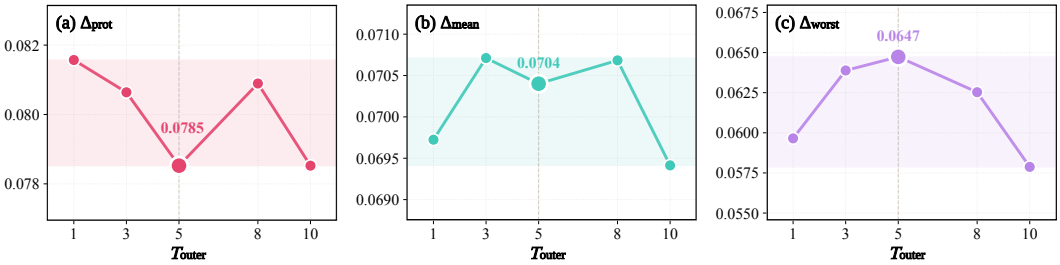

Figure A1: Outer-round sensitivity from existing NYC runs. Scores use the same descriptive A1–A3 aggregation as Table 1.

Table A1: Existing independent likelihood-diagnostic runs under the frozen trajectory LM. Higher mean log-likelihood indicates greater alignment with that scorer. Protected rows use three seeds; clean references use one run. These runs are not paired with the main-table protected samples.

Method	NYC ↑	TKY ↑
Clean	-3.8272	-3.9270
GHOST	-10.0208 ± 0.0132	-9.2804 ± 0.0586
PGD	-10.3664 ± 0.0261	-9.7827 ± 0.0803
EM	-10.1250 ± 0.0363	-8.6316 ± 0.0593
TS-UE	-10.0666 ± 0.0189	-8.5242 ± 0.0423

C Per-Method Failure Modes

TS-UE. The embedding-based adaptation suppresses direct-training accuracy but yields less favorable restored accuracy under lookup-table inversion. This pattern suggests that the projection can retain exploitable inverse correspondences. Determinism alone does not establish invertibility, and the observation concerns this discrete adaptation rather than time-series UE methods generally. Its lower substitution rate further complicates attribution to a single mechanism.

PGD. Deterministic adversarial candidate search provides strong direct-training suppression, showing that a frozen trajectory prior is not necessary for a large protection gap in this benchmark. Post-restoration accuracy is therefore important when assessing the added value of the prior and stochastic sampling.

EM. The high-support candidate rule provides weaker direct-training suppression in the reported comparison. A small survival gap alone does not establish strong protection: high protected accuracy leaves less room for improvement through restoration. Absolute restored accuracy and clean-referenced protection gaps should therefore be considered jointly.

GHOST. GHOST combines a competitive direct-training gap with favorable restored accuracy, particularly under A4. These results support the combined scoring-and-sampling design at the selected operating point. They do not isolate the contribution of the LM from stochastic selection or differing optimization budgets. The ablations characterize operating-point sensitivity without resolving these attribution limits.

Table A2: NYC score-weight sensitivity, seed 42 only. Each entry is recomputed from saved branch records using branch-specific baselines.

(α, β)	$\Delta_{\text{prot}} \uparrow$	$\Delta^{(2)}_{\text{surv}} \downarrow$	$\Delta^{(3)}_{\text{surv}} \downarrow$	$\Delta_{\text{mean}} \uparrow$
(1.0, 1.0)	0.0781	0.0091	0.0117	0.0712
(1.0, 1.5)	0.0729	0.0197	0.0113	0.0625
(1.5, 1.0)	0.0786	0.0144	0.0106	0.0716
(2.0, 0.5)	0.0810	0.0076	0.0146	0.0754

Table A3: NYC entropy-target sensitivity, seed 42 only, with score weights $(2.0, 0.5)$. Target entropy is capped by each candidate set's capacity.

Target η (bits)	$\Delta_{\text{prot}} \uparrow$	$\Delta_{\text{mean}} \uparrow$	$\Delta_{\text{worst}} \uparrow$
0	0.0837	0.0763	0.0691
1	0.0777	0.0712	0.0679
2	0.0731	0.0685	0.0641
3	0.0718	0.0688	0.0652

D Detailed Experimental Setup

D.1 Datasets and Preprocessing

We use Foursquare-NYC and Foursquare-TKY from the TSMC2014 release (Yang et al., 2015). Users and POIs with fewer than ten check-ins are filtered, and gaps longer than 24 hours separate sessions. Sessions are ordered temporally within each user. The reference preprocessing allocates the oldest $\lfloor 0.8n \rfloor$ sessions to training and the next $\lfloor 0.1n \rfloor$ to validation, with the remainder used for testing; per-user rounding changes the aggregate fractions. The reported processed splits contain 1,083 users and 5,135 POIs for NYC, with 10,958/937/2,359 train/validation/test sessions, and 2,293 users and 7,873 POIs for TKY, with 34,902/3,442/6,393 sessions. Vocabulary construction precedes the temporal split, so the setup is transductive in POI identities. Neither surrogate fitting nor LM fitting uses held-out sessions.

D.2 Architecture and Training

The architecture and optimization details below describe the reference implementation. The reported outcomes are reconstructed from saved runs; fields absent from those records and their historical code revisions cannot be retrospectively verified. The victim is a two-layer causal Transformer with 128-dimensional embeddings, four attention heads, and a maximum history length of 100. We refer to it as a Transformer victim rather than a reproduction of STAN (Luo et al., 2021). Victims are trained from scratch using AdamW, learning rate 10^{-3}, batch size 256, and clean-validation early stopping. Thus the benchmark grants access to clean validation feedback. The LM uses the same architecture, is trained on the training split, and is reused across runs. The recommended GHOST settings are those in Section 4.2. Restoration uses a default paired-session fraction of 5%; A2 is trained by masked cross-entropy and applied for three argmax passes.

D.3 Baseline Implementations

All methods share the geographic/category candidate construction and surrogate architecture. The label *PGD* denotes the existing deterministic error-maximizing candidate-search baseline, which selects the lowest clean-target logit; it is a discrete comparator inspired by adversarial optimization (Madry et al., 2018), not a continuous projected-gradient implementation. *EM* is a discrete error-minimizing-style baseline selecting high target support (Huang et al., 2021). *TS-UE* denotes a time-series-inspired baseline using embedding-space loss minimization followed by projection to POI candidates (Jiang et al., 2024); it is not a full reproduction of that paper's control-vector objective. The reference entry point uses three warm-up/inner epochs for PGD and EM, whereas GHOST uses five; TS-UE uses its separate embedding-optimization schedule. The available comparison is consequently not compute-matched. GHOST-Sym and GHOST-EF are internal sensitivity variants.

Algorithm 1 GHOST alternating trajectory perturbation

Input: clean sessions, candidate sets C, frozen LM T_ψ, weights, temperature, outer and inner budgets.
Warm up surrogate S_ϕ on clean history–target pairs.
Initialize each $\tilde{s} \leftarrow s$.
for $t = 1, \ldots, T_{\text{outer}}$ **do**
    for each batch of selected positions (s, i) **do**
        Keep positions without candidates unchanged.
        Evaluate Eqs. (3)–(5) on the current histories.
        Sample candidates with Eq. (6) and update the batch.
    end for
    Refit S_ϕ on $(\tilde{s}_{<i}, p_i)$ for E_{inner} epochs.
end for
return perturbed sessions $\{\tilde{s}\}$.

D.4 Evaluation Metrics

Let $c_{a,k}$ and $b_{a,k}$ denote the clean-training and protected-training accuracies of branch Aa at seed k, and let $r_{a,k}$ denote its restored accuracy for $a \in \{2, 3, 4\}$. All accuracies are evaluated on clean test data. Define outcomes $o_{1,k} = b_{1,k}$ and $o_{a,k} = r_{a,k}$ for $a > 1$, with branch-specific clean-referenced gaps $g_{a,k} = c_{a,k} - o_{a,k}$. Different branches may have different clean and protected realizations. We compute

$$\Delta_{\text{prot},k} = c_{1,k} - b_{1,k}, \qquad \Delta^{(a)}_{\text{surv},k} = r_{a,k} - b_{a,k}, \tag{10}$$

$$\Delta_{\text{mean},k} = \tfrac{1}{3} \sum_{a=1}^{3} g_{a,k}, \qquad \Delta_{\text{worst},k} = \min_{a \in \{1,2,3\}} g_{a,k}. \tag{11}$$

Baseline alignment. The survival gap uses the protected accuracy from the *same* restoration run. It therefore need not equal the difference between the displayed Aa and A1 means. Higher protection and aggregate gaps indicate lower predictive accuracy relative to the specified clean reference; lower survival gaps indicate less recovery relative to the corresponding protected baseline.

Aggregation and interpretation. Mean and worst gaps summarize A1–A3, with A4 reported separately. The worst gap is the minimum branch-specific clean-referenced gap; it reduces to clean accuracy minus maximum outcome accuracy only when the clean reference is shared. Matched seed labels do not imply identical protected releases across branches. These summaries are descriptive, not paired worst-case attack estimates or attack-optimal security bounds.

Reporting. All summaries are computed from saved per-run records using the stated branch references. We report means and sample standard deviations over three seeds where all three are available.

E Pseudocode

F Analysis of Manifold-Guided Substitution

Prior-guided adversarial selection. Equation (9) expresses the sampler as an adversarial tilt of a temperature-adjusted trajectory prior. The prior changes the relative preference among candidates, while the adversarial factor favors substitutions that challenge the surrogate. This identity characterizes candidate selection rather than establishing a privacy guarantee.

Denoising depends on the joint distribution. Let X and Y denote clean and protected POIs, with context h. Writing $Q(y \mid x, h)$ for the induced substitution channel, the population-optimal

predictive distribution under cross-entropy is

$$p(x \mid y, h) = \frac{p(x \mid h)Q(y \mid x, h)}{\sum_v p(v \mid h)Q(y \mid v, h)}. \tag{12}$$

The posterior depends on the coupling between clean and protected locations. Plausible outputs can therefore remain recoverable from paired examples. In particular, matching the clean output distribution does not imply an identity denoiser. The reported restored accuracies measure downstream training performance after the implemented restoration procedures, not Bayes-optimal reconstruction error.

Frequency-table inversion. A3 estimates a context-free version of Eq. (12), whereas A4 additionally conditions on the preceding protected token. Sampling can distribute an anchor's outputs across several candidates, and overlapping candidate sets can create multiple possible origins for a protected POI. These properties may complicate finite-sample inversion, but neither guarantees a uniform posterior: POI frequencies and candidate asymmetries also affect recoverability. The A3/A4 results therefore characterize the evaluated estimators at the specified leakage levels.

Distinct roles of the score components. The adversarial term alters the surrogate-facing prediction problem, while the LM term favors candidates according to learned mobility structure. Equation (7) characterizes their combination but does not establish that either component is individually necessary. A likelihood-only sampler may still disrupt training, and randomized adversarial selection may resist some restoration procedures. The available ablations compare operating points of the combined design rather than fully isolate these effects.

G Evaluation Scope and Deployment Considerations

Limitations. The results concern one Transformer victim and three specific restoration procedures. The neural denoiser is a BridgePure-inspired proxy, and cross-architecture transfer and stronger conditional generative attacks are untested. More fundamentally, A1 retains clean targets while restored training derives targets from reconstructed sessions. This limits causal interpretation of survival gaps and prevents the benchmark from establishing security for unrestricted complete-trajectory release. Publishing all overlapping clean-labeled prefixes could itself expose clean locations through linkage; that attack is not evaluated here. Candidate locality is adaptive rather than a hard 1 km bound, and the reported sequences do not satisfy a hard speed constraint. The LM likelihood diagnostic is internal, and non-target downstream utility is not measured. Baseline compute is not matched. These boundaries constrain the security and deployment conclusions, even when within-protocol rankings are favorable.

Deployment. GHOST specifies how a publisher can generate perturbed trajectories; it does not determine a safe release interface. A deployment must separately specify whether raw sessions, fixed training pairs, timestamps, and clean targets are exposed. Repeated releases also require separate analysis. Resampling from the same mechanism does not invalidate old clean–protected pairs, and linked outputs across releases can accumulate information. Our leakage sweeps therefore characterize single-release benchmark behavior rather than a renewable privacy budget. The frozen LM can be reused computationally, but this reuse has no implied security composition guarantee.

Extensions. The most direct extensions are to harmonize label construction across all branches, test prefix-linkage and stronger sequence restoration, and evaluate transfer across victim architectures. An independent mobility scorer and a specified non-target task would clarify whether likelihood guidance preserves useful structure beyond the optimized LM. These extensions follow from the separation between predictive disruption, trajectory compatibility, and inverse recoverability, which are distinct properties even when optimized within one framework.