BEYOND TRAJECTORY IMITATION: BEHAVIORALLY ALIGNED SYNTHETIC MOBILITY GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing synthetic mobility generators primarily optimize day-level trajectory realism, but realistic trajectories can still misrepresent how an individual moves. We study this missing objective as behaviorally aligned synthetic mobility generation: generating trajectories that realize an explicit multi-scale behavioral specification. We introduce *CTGTrace*, which represents individuals through 19 persistent characteristics and 16 daily intentions and realizes them through structured spatial generation, multi-day planning, and closed-loop trajectory decoding. Across five heterogeneous mobility populations and ten published generators, *CTGTrace* achieves a mean normalized request-realization error of 0.175 with correlation $r = 0.925$. We further show that the framework supports partial specifications, scenario transfer, and generation without an underlying real individual, and release the resulting synthetic dataset and code.

1 INTRODUCTION

Human mobility data supports a wide range of applications, including transportation planning, urban analytics, epidemic modelling, mobility prediction, and location-based services Gonzalez et al. (2008); Song et al. (2010b); Yabe et al. (2024b). However, obtaining large-scale, longitudinal mobility trajectories remains difficult. Direct collection requires sustained sensing and substantial processing, while many large-scale sources, such as mobile-network and application-derived traces, are owned by private providers and are available only commercially or under restricted access Mahajan et al. (2022); Yabe et al. (2024a). Privacy concerns further constrain the sharing of fine-grained individual trajectories De Montjoye et al. (2013). Synthetic mobility generation offers an alternative by producing artificial trajectories that preserve useful properties of real mobility without requiring the direct release of observed individual traces Zhu et al. (2023); Deng et al. (2025b); Ma et al. (2026).

Recent work has advanced synthetic mobility generation along several complementary directions. Early neural approaches such as MoveSim use adversarial sequence models to reproduce observed mobility transitions while incorporating urban spatial structure, while PateGAIL and PateGAIL++ formulate trajectory synthesis as imitation learning with privacy- and utility-aware training Feng et al. (2020); Wang et al. (2023); Ma et al. (2026). MBP-GAIL further conditions generation on discrete moving-behavior labels obtained by clustering POI-context transition patterns Lin et al. (2023), whereas MIRAGE combines a neural temporal point process with exploration and preferential-return dynamics Deng et al. (2025b). Diffusion-based methods instead model trajectory distributions directly, with extensions for road topology and trip attributes, user-profile conditioning, or collaborative mobility structure Zhu et al. (2023; 2024); Song et al. (2024); Zhang et al. (2025). Marionette broadens conditional generation to global and time-local contexts such as calendar, weather, and events Deng et al. (2025a). More recently, LLM-based approaches model personal mobility through semantic activity reasoning and mobility intentions: LLMob uses LLM agents to infer individual activity patterns and motivations, CoPB reasons over planned-behavior factors to generate subsequent activities, and ELLMob reconciles habitual mobility with event-driven constraints Wang et al. (2024); Shao et al. (2026); Wang et al. (2026). These methods substantially improve trajectory realism and conditional generation, but their controls are primarily trip attributes, profile/context variables, discrete behavior classes, or inferred intentions rather than an explicit multi-scale specification whose realization is evaluated for each generated individual.

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

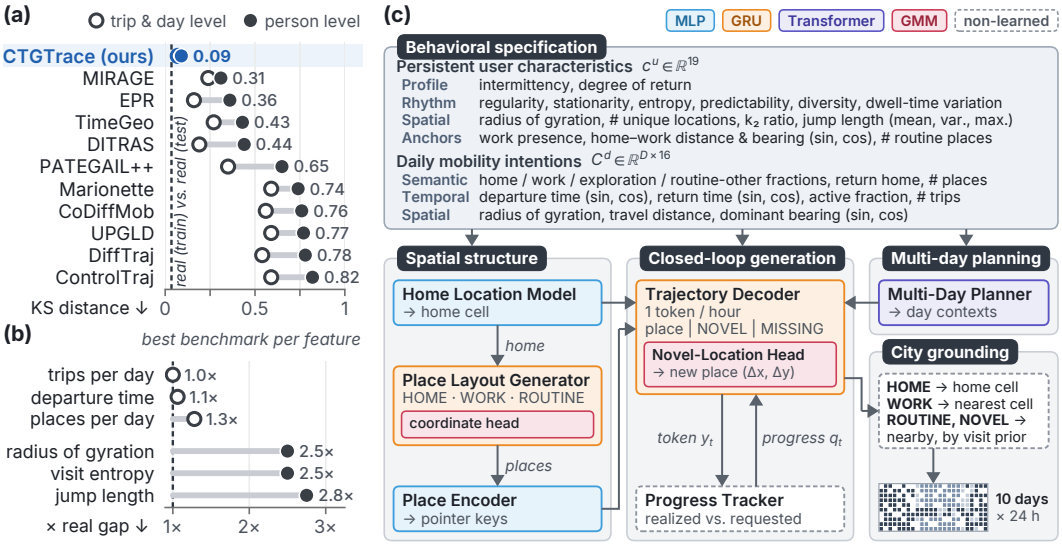

Figure 1: **Motivation and overview of *CTGTrace*.** (a) distance to real data for trip- and day-level statistics (○) and person-level characteristics (●), averaged over five populations. (b) Best benchmark on each feature, relative to the real (train) vs. real (test) gap (dashed). (c) Overview of *CTGTrace*.

Limitation of existing approaches. Despite this progress, existing methods largely treat synthetic mobility as trajectory imitation, possibly conditioned on trip attributes, contexts, or constraints. This overlooks that individuals differ systematically in how they move: how much they explore or return, how regular and predictable they are, and how far they range Gonzalez et al. (2008); Pappalardo et al. (2015); Alessandretti et al. (2018); Teixeira et al. (2021); Amichi et al. (2022); Esper et al. (2024); Félix et al. (2026b;a); exploration–return dynamics alone separate Routiners (few places, long returns), Regulars, and Scouters (many places, long explorations) Amichi et al. (2022). These differences are largely hidden in trip- and day-level statistics, which generators are usually evaluated on and reproduce comparatively well (Fig. 1(a)). Measured per person on 15 characteristics (leaving out work anchor-related characteristics), every published generator moves further from real data; even the closest lies at a KS distance of 0.31, against 0.04 between real training and test data. Even the best benchmark on each feature reproduces trips and places per day almost like real data, but remains 2.5–$2.8\times$ further from real on person level mobility features like radius of gyration, visit entropy, and jump length (Fig. 1(b)). Realistic trajectories and statistics therefore do not ensure realistic individuals.

Positioning. We characterize mobility behavior not by the specific locations and times in a trajectory, but through mobility metrics that capture how an individual moves, such as exploration–return tendency, regularity, and spatial range. We argue that synthetic mobility should therefore be behaviorally aligned: the mobility exhibited by a generated individual should be consistent with an explicit description of that individual's intended behavior. We formulate this description at two complementary scales. Persistent user characteristics describe behavior across the observation window, including exploration–return tendencies, mobility rhythm, spatial extent, and recurrent anchors, while daily mobility intentions specify how that behavior unfolds on individual days through activity allocation, departure and return structure, mobility extent, and travel patterns. Together, they form a multi-scale behavioral specification of 19 persistent characteristics and 16 daily intentions, grounded in established mobility-analysis literature Félix et al. (2026a); Amichi et al. (2022); Félix et al. (2026b). These dimensions capture complementary aspects of mobility rather than statistically independent or causally complete variables. Realizing such mobility behavioural specifications in generated trajectories is challenging because its properties emerge at different scales or aspects: recurrent anchors define persistent spatial structure, daily intentions govern individual days, and behaviors such as exploration, preferential return, regularity, and diversity emerge only over the trajectory as a whole. A generator must therefore coordinate decisions across temporal and spatial dimensions while preserving these mobility behaviors when mapping generated trajectories onto a real city.

Contributions. To address these challenges, we introduce *CTGTrace*, a structured generative framework for behaviorally aligned synthetic human mobility. Rather than asking a single sequence model to implicitly realize all behavioral properties, *CTGTrace* decomposes generation into multiple

properties that define user behavior including home (user origin) selection, recurrent-place generation, multi-day behavioral planning, causal trajectory realization, and city map grounding. The framework combines learned generative components with explicit geometric transformations and trajectory-level behavioral state. This design allows stochastic mobility patterns to be learned from data while exposing information about trajectory-wide behavioral progress during generation so that properties operating across different scales can be realized coherently rather than emerging only as an incidental consequence of local next-step predictions. Our contributions are:

- **Behaviorally aligned synthetic mobility generation.** We formulate mobility generation around an explicit multi-scale description of how an individual is intended to behave, combining persistent user characteristics with daily mobility intentions. We realize this formulation through *CTGTrace*, which coordinates spatial structure, multi-day planning, and closed-loop trajectory generation.
- **Flexible multi-scale behavioral specification.** *CTGTrace* flexibly accepts persistent, partial, or complete behavioral specifications, optionally combined with known spatial evidence, and generates completetrajectories while inferring unspecified behavioral aspects. It can also synthesize complete individuals without requiring behavioral controls from an existing real person.
- **Comprehensive evaluation of synthetic mobility generation.** We evaluate generation at two complementary levels: individual-level realization of the requested behavior, and population-level quality across behavioral and spatial fidelity, independent sequence realism, downstream utility, diversity and non-imitation, and privacy. Our study spans five heterogeneous mobility populations and ten published benchmark generators.
- **Synthetic mobility dataset and tool release.** 1.16 million synthetic people across five city populations, ten days each at one-hour resolution, in three mobility profiles: Routiner, Regular and Scouter. A next-location prediction model trained only on synthetic people predicts real test people's next location with up to 87% Acc@1, exceeding training on real people. For another temporal resolution or observation duration, trajectories can be regenerated using the released code. Both code and synthetic data are available here[1].

2 PRELIMINARIES

We represent the mobility of an individual user u over D days and H time intervals per day as

$$\mathbf{X}_u = (x_{u,1}, \ldots, x_{u,T}), \qquad T = DH, \tag{1}$$

where $x_{u,t}$ denotes a location in the map of the target city, or a missing observation. In the main setting, observation window $D = 10$ and $H = 24$, yielding $T = 240$ hourly records per individual.

Behavioral specification - In this work, we capture behavior in mobility through the set of characteristics extracted from the $\mathbf{X}_u$, as detailed in Appendix A. We describe the intended individual user behavior at two complementary levels. *Persistent user characteristics* $\mathbf{c}^u \in \mathbb{R}^{19}$ describe behavior across the complete observation window, whereas *daily mobility intentions* $\mathbf{C}^d \in \mathbb{R}^{D\times 16}$ describe how mobility is intended to unfold on individual days. A binary mask $\mathbf{M}^d \in \{0,1\}^{D\times 16}$ indicates which daily intentions are supplied, giving

$$\mathcal{B} = \left(\mathbf{c}^u, \mathbf{C}^d, \mathbf{M}^d\right) \tag{2}$$

This specification spans the behavioral groups summarized in Fig. 1(c); the complete 35 dimensions and their definitions are provided in Appendix A.1. The dimensions describe complementary aspects of mobility behavior rather than statistically independent or causally complete variables. We also include work anchoring in the behavioral specification because work-attending behavior influences both daily mobility intentions and persistent mobility characteristics. We further categorize individuals as *Routiners*, *Regulars*, or *Scouters* based on exploration–return preferences Amichi et al. (2022); Félix et al. (2026b): Routiners repeatedly visit a small set of familiar places, Scouters explore more new locations, and Regulars lie between these extremes.

Behavior versus spatial evidence - The behavioral specification describes how the individual is intended to move; optional spatial evidence instead identifies known recurrent places for that individual. These inputs are independent: daily intentions and spatial evidence can each be complete, partial, or absent. For fully de-novo generation, where no real individual provides $\mathbf{c}^u$, we sample a new persistent-characteristic vector from a Gaussian copula fitted to the training population and generate the remaining daily and spatial structure through the same pipeline.

[1] https://anonymous.4open.science/r/CTGTrace-482E

3 METHODOLOGY

We represent behavioral specifications using mobility metrics because they provide interpretable, directly measurable descriptions of how an individual moves, rather than the specific locations and times appearing in a trajectory. The selected characteristics capture complementary aspects of mobility at persistent and daily scales—including exploration–return preference, mobility rhythm, spatial extent, recurrent anchors, activity allocation, and temporal structure—so their realization requires coordination across different spatial and temporal scales. *CTGTrace* therefore separates generation into spatial-structure construction, closed-loop generation, multi-day planning, and city grounding, as illustrated in Fig. 1(c).

3.1 CONSTRUCTING THE PERSISTENT SPATIAL STRUCTURE

Home selection - Home is selected before both spatial-layout and temporal generation. This is important because home is not merely the origin of a local coordinate system: residential context based on home selection can influence both the feasible organization of recurrent destinations and how those destinations are used over time. A separately trained Home Location Model scores plausible residential cells conditioned on the persistent user characteristics. The selected home is then provided independently to both the Place Layout Generator and the Trajectory Decoder; if home is supplied as spatial evidence, the observed location is retained instead.

Place layout generator - Human mobility typically alternates between repeated visits to a small repertoire of familiar locations and less frequent exploration. We therefore generate recurrent and novel places separately. The recurrent scaffold contains HOME, available WORK anchors, and remaining ROUTINE places; each place has a home-relative coordinate, expected visit count, and semantic role. One-off destinations are generated later when the trajectory decoder chooses to explore.

A key design choice is to separate the stochastic shape of this recurrent-place configuration from geometry already specified by the intended behavior. For example, home–work bearing determines orientation, while mobility range largely determines scale of movement. Rather than asking a neural model to relearn these transformations, the Place Layout Generator learns a canonical home-relative configuration and the requested frame is imposed explicitly:

$$\mathbf{l}_p = s_u \, \mathbf{R}(\theta_u) \, \hat{\mathbf{l}}_p, \tag{3}$$

where $\hat{\mathbf{l}}_p$ is a generated canonical location, $\mathbf{R}(\theta_u)$ applies the intended orientation, and s_u restores the intended spatial scale. Discrete structural characteristics such as work presence and the number of routine places are likewise instantiated explicitly. When spatial evidence is available, known places remain fixed and only the missing structure is synthesized. The exact canonicalization and anchor-preserving scale adjustment are given in Appendix A.3.

3.2 BEHAVIORAL PLANNING AND CLOSED-LOOP GENERATION

From global characteristics to trackable targets - Several persistent characteristics cannot be attached to a single time slot. Exploration, return, regularity, stationarity, and repertoire size describe statistics of the trajectory accumulated over time. Static conditioning alone therefore gives an autoregressive decoder no indication of whether it is currently under- or over-realizing these characteristics. Before decoding, *CTGTrace* converts compatible persistent characteristics into compact trajectory-level behavioral targets, such as the intended number of observed events and distinct places, exploration–return structure, and repeated occupancy. These are deterministic consequences of the behavioral specification rather than predictions of another neural network. The Multi-Day Planner, implemented as a 4-layer Transformer, combines these targets with the persistent characteristics, available daily intentions, their masks, and temporal context to produce one representation for each day. Unspecified daily intentions are not imputed before generation; they remain free for the trajectory decoder to realize.

Causal trajectory realization - A 3-layer GRU generates a single causal sequence across all 240 hours. At each hour it emits one of the recurrent-place identities, NOVEL, or MISSING. Recurrent places are represented by MLP-derived keys and selected through pointer-style scoring, allowing place desirability to depend on visitation history without changing place identity. When NOVEL is selected, an 8-component Gaussian-mixture head proposes a new home-relative destination.

The central mechanism is a feedback loop between the generated trajectory and the behavioral specification. After every decision, a deterministic Progress Tracker updates causal statistics of the trajectory prefix—including observations, distinct places, exploration and return behavior, runs, and dwell structure—and computes what remains to be realized. For example, the remaining exploration demand is represented as

$$q_t^E = \frac{E^\star - E_t}{\max(1, N^\star - N_t)}, \tag{4}$$

where $E^\star$ and $N^\star$ are the target numbers of explorations and observations, while E_t and N_t denote what the current trajectory prefix has already produced. Analogous residual signals are maintained for the remaining trajectory-level and supplied daily characteristics. Thus, generation operates as:

trajectory prefix $\rightarrow$ behavioral progress $\rightarrow$ remaining demand $\rightarrow$ next decision

This feedback is what allows trajectory-level characteristics to influence local decisions throughout generation instead of appearing only as static conditioning variables. Semantic mobility states are kept consistent with the generated trajectory by construction. HOME and WORK are determined by the selected recurrent place, while EXPLORE is assigned when a previously unseen non-seeded place is first visited; otherwise the visit is treated as OTHER-ROUTINE. We therefore do not learn a separate state head that could contradict the generated location sequence.

3.3 City Grounding and Model Training

City-aware grounding - The generator operates over abstract place identities and home-relative coordinates, whereas its output must occupy real cells of the target city. We therefore separate which behavioral place is visited from which physical city cell realizes that place. Home is grounded first, followed by work anchors so that the intended anchor geometry is preserved. For routine and novel places, nearest-cell projection alone is insufficient: nearby urban cells can have very different visitation frequencies. Grounding therefore balances geometric proximity with an empirical destination-attractiveness prior estimated from the training population,

$$p(c \mid \mathbf{q}) \propto \exp\left[\lambda \log P_{\text{visit}}(c) - \frac{d(\mathbf{q}, c)}{\tau}\right], \tag{5}$$

where $\mathbf{q}$ is the generated coordinate. This preserves the intended geometry while avoiding unrealistic use of the city's spatial support.

Training - The Home Location Model is trained separately with a contrastive ranking objective. The remaining neural components are optimized jointly using

$$\mathcal{L} = \mathcal{L}_{\text{token}} + \mathcal{L}_{\text{exist}} + 0.15\,\mathcal{L}_{\text{novel}} + 0.15\,\mathcal{L}_{\text{layout}} + 0.15\,\mathcal{L}_{\text{count}} + 0.05\,\mathcal{L}_{\text{plan}}, \tag{6}$$

where $\mathcal{L}_{\text{token}}$ is cross-entropy over hourly location tokens, $\mathcal{L}_{\text{exist}}$ is binary cross-entropy for recurrent-place existence, $\mathcal{L}_{\text{novel}}$ and $\mathcal{L}_{\text{layout}}$ are Gaussian-mixture negative log-likelihoods for novel and recurrent-place coordinates, $\mathcal{L}_{\text{count}}$ supervises recurrent-place visitation counts, and $\mathcal{L}_{\text{plan}}$ is an auxiliary objective over masked daily intentions. During training, daily intentions are masked under multiple patterns so that one checkpoint supports complete, partial, and persistent-only specifications. Full optimization and implementation details are given in Appendix A.

4 Experimental Setup

Datasets. We evaluate on five mobility populations chosen to expose complementary sources of variation. Shanghai is a non-public, anonymized telecom-derived mobility dataset covering ten days, while Shenzhen is a public anonymized cellular dataset Zhang et al. (2015); together they provide geographic variation across two Chinese cities with substantially different observation densities. YJMob-A and YJMob-B are different cohorts from the same Japanese metropolitan region under business-as-usual mobility, while YJMob-B and YJMob-C use the same source cohort under business-as-usual and emergency conditions, respectively Yabe et al. (2024b). All main experiments use a common representation of ten days with 24 hourly slots per day, with person-disjoint training, validation, and test splits. Dataset construction, population sizes, spatial resolutions, missingness, and preprocessing are detailed in Appendix B.

Table 1: **Individual-level behavioral realization.** The first block separates persistent and daily realization. The mechanism block reports MAE/r for analytically imposed (A), closed-loop targeted (C), and emergent/learned (E) factors. Profile agreement summarizes the exploration–return profile. The final block compares *CTGTrace* with requester-specific optimized EPR on the four requested quantities expressible by EPR. All-35 values are mean±SD over five generation draws.

	Multi-scale realization			Realization mechanism			Profile		Matched Request-EPR	
Population	Pers.-19 ↓	Daily-16 ↓	All-35 ↓	A: MAE/r	C: MAE/r	E: MAE/r	Acc. ↑	κ ↑	Ours-4 ↓	EPR-4 ↓
Shanghai	.155	.077	.119±.003	.183/.921	.082/.975	.119/.959	.949	.912	.101	.444
Shenzhen	.185	.140	.165±.001	.196/.905	.124/.964	.177/.925	.921	.865	.138	.479
YJMob-A	.233	.124	.183±.002	.295/.852	.137/.951	.172/.924	.935	.889	.167	.505
YJMob-B	.265	.133	.205±.002	.350/.827	.153/.941	.186/.917	.917	.859	.177	.488
YJMob-C	.260	.133	.202±.002	.357/.814	.147/.946	.182/.917	.928	.882	.169	.486
Mean	.220	.121	**.175**	.276/.864	**.129/.955**	.167/.929	**.930**	**.881**	**.150**	.480

Generation settings. Our primary behavioral-alignment setting supplies the 19 persistent user characteristics, but no daily mobility intentions and no real place evidence. We additionally consider a *full* setting in which all daily intentions are supplied, and a *de-novo* setting in which no real individual provides the request.

Baselines. We compare with ten published trajectory generators spanning neural event models, diffusion, imitation learning, and classical mechanistic mobility models: MIRAGE Deng et al. (2025b), Marionette Deng et al. (2025a), PATEGAIL++ Ma et al. (2026), DiffTraj Zhu et al. (2023), ControlTraj Zhu et al. (2024), UPGLD Song et al. (2024), CoDiffMob Zhang et al. (2025), DITRAS Pappalardo & Simini (2018), TimeGeo Jiang et al. (2016), and EPR Song et al. (2010a). Each method is ported to the same 10×24 city-cell representation and evaluated with the same metric implementation on the same five populations. The common cross-method benchmark uses $n = 1{,}000$ generated individuals per population. Full reproduction details and all required adaptations are documented in Appendix C.

Evaluation protocol. We distinguish paired evaluation, which asks whether a generated individual realizes a supplied behavioral specification, from unpaired evaluation, which asks whether a generated population preserves properties of real mobility and therefore applies to every generator. Our evaluation covers individual behavioral alignment, population feature fidelity, independent sequence realism, conventional mobility distributions, population spatial structure, downstream train-on-synthetic/test-on-real utility, diversity and non-imitation, and empirical release privacy. Distributional and realism scores are calibrated against matched real-versus-real references wherever appropriate. For utility, every downstream model is tested exclusively on the real test set and compared with the same model trained on a size-matched real corpus. Metric definitions, null construction, and downstream task details are provided in Appendix D.

5 Results

We evaluate *CTGTrace* at two complementary levels. At the individual level, we test whether each generated trajectory realizes its supplied behavioral specification. At the population level, we evaluate the resulting synthetic release across behavioral and spatial fidelity, independent sequence realism, downstream utility, diversity and non-imitation, and privacy. We then isolate the contributions of individual components through ablations and test robustness beyond the main setting.

5.1 Individual-Level Behavioral Realization

Table 1 summarizes the results. The 19 persistent characteristics are evaluated under persistent-only generation, while the 16 daily intentions are evaluated when explicitly supplied; their merged score therefore evaluates each quantity only when requested. MAE is normalized by the training-population standard deviation. For further analysis, we group factors by how the request influences generation: analytically imposed (A), where the request directly determines generated structure; closed-loop targeted (C), where realization is explicitly tracked during decoding; and emergent/learned (E), where the request conditions the generator without an explicit realization rule.

***CTGTrace* realizes requested behavior at both scales.** Generated behavior tracks the request across all 35 dimensions (pooled $r = 0.925$) with a mean error of 0.175 SD, under a fifth of the between-person spread. Daily intentions average 0.121 SD, and even the more difficult persistent characteristics, which must emerge over ten days from the request alone, average 0.220 SD.

Alignment is not an artifact of imposed structure. The six imposed dimensions are the hardest to realize, since imposed structure must survive decoding and city grounding: an imposed routine location counts only if the trajectory visits it often enough to be recovered. Emergent dimensions, which no rule targets, come close to closed-loop ones, and even the seven with neither imposed structure nor residual-demand control match the closed-loop average (0.128 vs. 0.129 SD; Appendix E).

***CTGTrace* preserves requested exploration–return profiles.** The profile (Routiner, Regular, or Scouter) is recovered for over 91% of individuals (Cohen's $\kappa > 0.85$), as both defining quantities are realized closely (MAE/r of 0.075/0.964 for intermittency and 0.122/0.983 for degree of return). Since it depends only on these two, it complements rather than replaces the full evaluation.

***CTGTrace* follows requests far more closely than request-calibrated EPR.** As most generators lack a request interface, we optimize EPR's five person-specific parameters from each request and score a fresh trajectory on the four quantities EPR can express. Even with this per-person calibration, EPR's error is 2.7–$4.4\times$ that of *CTGTrace* in every population. Since both receive the same request, the advantage comes from how *CTGTrace* realizes it, not from extra information (Appendix E.3).

Structured generation, not conditioning alone, drives realization. A flat conditional Transformer (FCT) with the same inputs, predicting cells directly without the recurrent-place scaffold, Multi-Day Planner, pointer mechanism, Progress Tracker, or city grounding, has 3.0–$10.6\times$ *CTGTrace*'s persistent-request MAE on Shanghai, Shenzhen, and YJMob-B. Adding *CTGTrace*'s progress representation closes only 22–48% of this gap, so its other stages supply most of the gain (Appendix E.4).

5.2 Population-Level Synthetic Mobility Quality

Having established paired behavioral realization, we next evaluate whether the resulting synthetic population preserves the broader properties required of useful mobility data. Table 2 compares *CTGTrace* with ten published generators across complementary population-level axes. Values are averaged across five populations. The *CTGTrace* row uses fully synthetic individuals whose persistent characteristics are sampled from a Gaussian-copula approximation fitted to the training-population distribution, so no real individual underlies a released trajectory.

Population fidelity. ***CTGTrace* is by far the closest to real populations.** Feature KS measures, for each behavioral characteristic, the largest gap between the synthetic and real cumulative distributions, normalized so that two real samples score about $1\times$. Over the 35 characteristics, *CTGTrace* reaches $2.05\times$, against $6.59\times$ for the strongest baseline, MIRAGE, and 9.40–$10.92\times$ for the mechanistic references (Fig. 2; Appendix E.2). The person-level gap identified in the introduction thus largely closes, even though no real individual underlies these trajectories. Distribution JS, the mean Jensen–Shannon distance over thirteen conventional mobility distributions, agrees: *CTGTrace* reaches 0.107, close to the real–real 0.078 and under half of MIRAGE's 0.249, and the ranking holds under Hellinger, Wasserstein/IQR, and KS (Appendix F.1).

Sequence realism. ***CTGTrace* comes closest to real temporal structure.** Realism averages ten diagnostics of properties the specification does not directly summarize (motifs, dwell times, semantic transitions, switching rates, circadian patterns), each normalized by real–real variability. *CTGTrace* scores $1.42\times$ the real-data null, against $3.58\times$ for MIRAGE and 5.60–$6.90\times$ for the mechanistic references, and given a request it reaches the real-data level (0.93–$0.99\times$ vs. $0.90\times$). A learned two-sample classifier separates the primary baselines from real data almost perfectly (AUC 0.993) but *CTGTrace* only weakly (0.587), and *CTGTrace*'s speed tail is also the smallest (Appendix F.2).

Spatial fidelity. ***CTGTrace* reproduces origin–destination flows most closely.** Cell JSD compares where visits fall across city cells; OD JSD compares how trips spread over origin–destination cell pairs. *CTGTrace* has the lowest OD JSD of all generators (0.2012 vs. 0.1724 between real samples), ahead of EPR, DITRAS, and TimeGeo, which are built around aggregate destination laws. Its cell JSD matches the real–real reference (0.3075 vs. 0.3055), with MIRAGE slightly lower, and it best preserves the ranked popularity of places, staying competitive on other spatial views (Appendix F.3).

Independent utility. **Models trained on *CTGTrace* nearly match models trained on real data.** Each model is trained on the release and tested on real data, on targets outside the specification. For next-visit prediction, Acc@1 is the share of real visits whose top prediction is correct: *CTGTrace* reaches 0.542, against 0.548 with real training and 0.515 for MIRAGE. For 256-zone occupancy nowcasting, R^2 is the share of variance in real zone populations explained: *CTGTrace* and MIRAGE tie at 0.91 (real: 0.929), while all others stay below 0.80. For an SIR epidemic calibrated on real

Table 2: **Population-level synthetic mobility evaluation across five populations.** Feature KS and Dist. JS measure behavioral and conventional distributional fidelity; Realism summarizes the independent sequence battery. Coverage and train similarity are calibrated against equal-size real references and are best near one; MIA is best near 0.5.

	Behavioral fidelity		Seq. realism	Spatial fidelity		Independent utility			Behavior preservation			Diversity / non-imitation		Leakage check
Method	Feat. KS ↓	Dist. JS ↓	Realism ↓	Cell JSD ↓	OD JSD ↓	Next Acc. ↑	Occ. R^2 ↑	SIR MAPE ↓	Prefix F1 ↑	Rec. NDCG ↑	Act. F1 ↑	Cover. → 1	Train sim. → 1	MIA → .5
Real reference	*1.00*	*.078*	*.90*	*.3055*	*.1724*	*.548*	*.929*	*12.1*	*.715*	*.900*	*.441*	*1.00*	*1.00*	*≈.50*
***CTGTrace*, de-novo**	**2.05**	**.107**	**1.42**	.3075	**.2012**	**.542**	**.909**	**21.1**	**.696**	**.895**	**.450**	**.89**	**.86**	.503
MIRAGE [KDD'25]	6.59	.249	3.58	**.2904**	.3969	.515	.908	24.4	.542	.810	.375	.81	.61	.502
DITRAS [DMKD'18]	10.92	.377	5.60	.3100	.2378	.362	.792	221.8	.196	.736	.207	.75	.73	.504
TimeGeo [PNAS'16]	9.40	.382	6.90	.4359	.3637	.448	.785	136.2	.443	.875	.242	1.30	.62	.513
EPR [Nat. Phy.'10]	10.41	.325	6.75	.4121	.2086	.450	.785	134.3	.303	.787	.205	1.29	.65	.504
PATEGAIL++ [ICLR'26]	16.04	.630	12.4	.4253	.3015	.354	.773	110.8	.173	.711	.153	1.29	.43	.500
DiffTraj [NeurIPS'23]	17.48	.664	10.9	.9899	.9616	.388	.502	709.4	.206	.772	.162	.11	.29	.488
CoDiffMob [WWW'25]	17.82	.739	13.1	.8398	.8280	.367	.556	449.8	.409	.850	.126	.87	.19	.493
Marionette [KDD'25]	17.83	.676	13.1	.8725	.8961	.325	.547	535.6	.240	.648	.239	.09	.81	.493
UPGLD [TKDD'24]	18.25	.707	11.5	.9923	.9943	.319	.481	804.0	.123	.616	.129	.01	.37	.498
ControlTraj [KDD'24]	19.46	.736	12.4	.9932	.9581	.292	.434	622.3	.126	.724	.140	.05	.23	.492

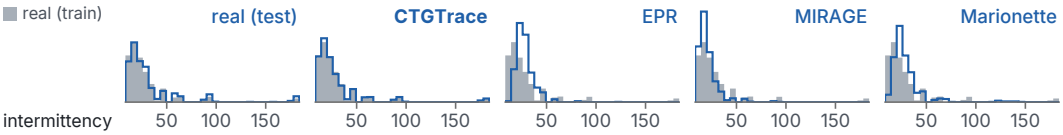

Figure 2: Intermittency distributions of synthetic populations compared with the real test population.

mobility, MAPE is the relative error between curves simulated on synthetic and real co-location: *CTGTrace* is lowest among generators (21.1% vs. 24.4% for MIRAGE and over 110% for the rest), though large real–real variation makes this a stress test, not a fine-grained ranking (Appendix F.4).

Behavior preservation. **Behavioral structure remains learnable from the release.** Models trained on *CTGTrace* nearly match real training at predicting a person's mobility profile from the first three days (Prefix F1 0.696 vs. 0.715) and at recommending their secondary places, scored by top-3 ranking quality (NDCG@3 0.895 vs. 0.900); they slightly exceed it on activity-state prediction (Act. F1 0.450 vs. 0.441) and lead every other generator on all three. As these targets overlap the specification, they measure preserved behavioral structure, not independent utility.

Diversity and non-imitation. ***CTGTrace* neither copies training individuals nor collapses.** No generated trajectory is an exact ordered copy of a training trajectory, whereas the same detector flags 84–99% of a retrieval control's copies of training data. Train similarity, the maximum Jaccard overlap between a person's visited cells and any training individual's, relative to held-out real people, is 0.86: generated people are no closer to the training set than unseen real ones. Coverage, the distinct cells reached relative to an equal-size real population, is 0.89. *CTGTrace* is closest to 1 on both, while some generators over-disperse (up to 1.30) or collapse spatially (down to 0.01; Appendix F.5).

Privacy. **Fidelity comes without detectable membership leakage.** Membership inference (MIA) tests whether an attacker can tell that a given real trajectory was used for training, with ROC-AUC 0.5 at chance. The strongest attack we evaluate reaches only 0.503 on the fully synthetic release. This holds for releases not tied to real individuals (requester-conditioned ones retain more person-specific signal) and is an empirical audit, not a formal or checkpoint-level guarantee (Appendix F.6). We report this as a sanity check against memorization rather than as a privacy guarantee; formal guarantees (e.g., differentially private training) and stronger, mobility-specific attacks are beyond the scope of this work, which focuses on generation quality.

5.3 Ablation Study

The ablations reveal a clear hierarchy: recurrent-place usage provides the dominant contribution to behavioral realization, trajectory-level progress gives a smaller consistent gain, and the remaining spatial components refine the result.

Recurrent-place usage is the behavioral backbone. Removing visitation-count supervision increases behavioral error by 0.588–0.686 SD across the three tested populations, far more than any other intervention. The checkpoints remain active, but generated trajectories no longer use recurrent places with the intended frequencies, showing that plausible recurrent locations alone are insufficient. The decoder also receives recurrent-place counts inside its pointer keys. Although training uses observed counts while generation uses place layout generator's predictions, this asymmetry does

Table 3: **Behavioral ablations.** Mean increase in released-grid MAE after removing or replacing each component; positive values are worse. Retrained interventions use three seeds per arm.

Intervention	Populations	Mean $\Delta\text{MAE}_{\text{SD}} \uparrow$
Remove recurrent-place count supervision	SH/SZ/YJ-B	**+0.597**
Remove coordinate-mixture objective	SH/SZ/YJ-B	+0.020
Remove trajectory progress information	SH/SZ/YJ-B	+0.117
Canonical layout → direct layout	SZ/YJ-B	+0.009
Remove daily context from layout	SZ/YJ-B	+0.008

not explain the result: removing the count channel or replacing the training input with place layout generator's own prediction increases combined behavioral error by 18–27%. In contrast, removing count supervision increases error by 3.5–$5.8\times$ (Appendix G). Thus, recurrent-place count supervision, rather than the train–generation count mismatch, is the critical mechanism.

Progress-aware generation provides a complementary gain. Removing trajectory-level progress information increases MAE by 0.110–0.126 SD on all three populations and reduces profile agreement. A decode-time stress test produces a larger failure when person-specific progress is replaced by population-typical progress, but because this is out of distribution, the retrained ablation provides the formal effect size.

Other components refine rather than explain alignment. Layout realization has a measurable effect, while work-existence, state-deficit, routine-count, and novel-jump corrections are small or population dependent. Together, the ablations indicate that learned recurrent-place usage and progress-aware decoding perform the main behavioral work rather than late deterministic corrections (Appendix G).

City-aware grounding addresses population placement. Home-density and destination-attractiveness grounding mainly improve where synthetic individuals are placed in the city, substantially improving occupancy utility while leaving next-visit prediction nearly unchanged. The intervention incurs a small behavioral-fidelity cost on some populations and is therefore treated as a spatial fidelity–utility trade-off rather than a source of behavioral alignment (Appendix H).

5.4 Robustness Beyond the Main Setting

The conclusions remain stable under partial behavioral specifications, alternative temporal resolution, and longer observation windows. Providing more persistent characteristics progressively improves profile agreement; using YJMob's native 48 half-hour slots provides no consistent profile benefit and increases fine-grained behavioral error; and extending the horizon from ten to fifteen days produces only modest changes. Under YJMob-B→YJMob-C transfer, evaluated on target users disjoint from the training population, the YJMob-B model reaches feature KS 1.518 against 1.560 for the matched YJMob-C model. Spatial fidelity and downstream utility remain comparable, while the generated profile distribution closely matches the emergency-period population. This demonstrates transfer across mobility scenarios within the same dataset family without retraining (Appendix I.4).

6 Conclusion

We introduced behaviorally aligned synthetic human mobility generation, shifting the goal from producing plausible trajectories alone to generating individuals whose mobility reflects an intended behavioral description. *CTGTrace* realizes this through a structured combination of spatial scaffold generation, multi-day planning, closed-loop trajectory realization, and city grounding. Across five heterogeneous populations and ten published generators, the resulting trajectories preserve strong individual behavioral alignment while remaining realistic and useful under independent evaluations. More broadly, our results suggest that synthetic mobility should be judged not only by whether generated trajectories look real, but by whether the generated individuals behave as intended.

7 Ethics Statement

Human mobility traces are privacy-sensitive and may reveal individual behavioral patterns. All source datasets used in this work contain anonymized mobility records, and we make no attempt to identify or re-identify individuals. Shenzhen and YJMob are publicly available anonymized datasets, while Shanghai is a non-public, anonymized telecom-derived dataset used solely for research under its access conditions. We do not release any user trajectories of the Shanghai dataset; however, an anonymised synthetic dataset contains only trajectories generated by *CTGTrace*. Our membership-inference evaluation provides an empirical privacy audit rather than a formal privacy guarantee. We adhere to the ICLR Code of Ethics.

8 Reproducibility Statement

We provide an anonymous implementation and the generated synthetic dataset with the submission. Appendix A specifies the complete architecture, objectives, and generation procedure; Appendix B describes datasets and preprocessing; Appendix C documents benchmark reproduction and adaptations; and Appendix D defines the evaluation protocol.

9 AI Use Statement

Generative AI tools were used to provide feedback on the written manuscript and to assist with visual refinement of figures. All AI-assisted suggestions and outputs were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

References

Laura Alessandretti, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli. Evidence for a conserved quantity in human mobility. Nature human behaviour, 2(7):485–491, 2018.

Licia Amichi, Aline Viana Carneiro, Mark Crovella, and Antonio Loureiro. Revealing an inherently limiting factor in human mobility prediction. IEEE Transactions on Emerging Topics in Computing, 11(3):635–649, 2022.

Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3(1):1376, 2013.

Bangchao Deng, Ling Ding, Lianhua Ji, Chunhua Chen, Xin Jing, Bingqing Qu, and Dingqi Yang. Marionette: Fine-grained conditional generative modeling of spatiotemporal human trajectory data beyond imitation. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 463–473, 2025a.

Bangchao Deng, Xin Jing, Tianyue Yang, Bingqing Qu, Dingqi Yang, and Philippe Cudre-Mauroux. Revisiting synthetic human trajectories: Imitative generation and benchmarks beyond datasaurus. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1, pp. 201–212, 2025b.

João Paulo Esper, Aline Carneiro Viana, and Jussara M Almeida. Beauty or beast: Human behavioral insights and learning power of federated mobility prediction. In Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, pp. 325–337, 2024.

Lucas GS Félix, Anne Josiane Kouam, Aline C Viana, Nadjib Achir, and Jussara M Almeida. How behavioral exposure reveals privacy risks beyond location uniqueness. In ESORICS 2026-31st European Symposium on Research in Computer Security, 2026a.

Lucas GS Félix, Anne Josiane Kouam, Aline Carneiro Viana, Nadjib Achir, and Jussara M Almeida. Tracing the unseen: a deep dive into mobility behavior exposure. EPJ Data Science, 2026b.

Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. Learning to simulate human mobility. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 3426–3433, 2020.

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

Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C González. The timegeo modeling framework for urban mobility without travel surveys. Proceedings of the National Academy of Sciences, 113(37):E5370–E5378, 2016.

Haowen Lin, Sina Shaham, Yao-Yi Chiang, and Cyrus Shahabi. Generating realistic and representative trajectories with mobility behavior clustering. In Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, pp. 1–4, 2023.

Yingjie Ma, Bijal Bharadva, Xin Zhang, and Joann Qiongna Chen. Pategail++: Utility optimized private trajectory generation with imitation learning. In The Fourteenth International Conference on Learning Representations, 2026.

Vishal Mahajan, Nico Kuehnel, Aikaterini Intzevidou, Guido Cantelmo, Rolf Moeckel, and Constantinos Antoniou. Data to the people: a review of public and proprietary data for transport models. Transport reviews, 42(4):415–440, 2022.

Luca Pappalardo and Filippo Simini. Data-driven generation of spatio-temporal routines in human mobility. Data Mining and Knowledge Discovery, 32(3):787–829, 2018.

Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. Nature communications, 6(1):8166, 2015.

Chenyang Shao, Bingbing Fan, Jingtao Ding, Yuan Yuan, Meng Wang, and Fengli Xu. Generating realistic human mobility data with hybrid large language model agent. In Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1, pp. 2758–2769, 2026.

Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. Nature physics, 6(10):818–823, 2010a.

Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. Science, 327(5968):1018–1021, 2010b.

Yiwen Song, Jingtao Ding, Jian Yuan, Qingmin Liao, and Yong Li. Controllable human trajectory generation using profile-guided latent diffusion. ACM Transactions on Knowledge Discovery from Data, 19(1):1–25, 2024.

Douglas do Couto Teixeira, Jussara M Almeida, and Aline Carneiro Viana. On estimating the predictability of human mobility: the role of routine. EPJ Data Science, 10(1):49, 2021.

Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection. In Proceedings of the 26th International Conference on Artificial Intelligence and Statistics, volume 206 of Proceedings of Machine Learning Research, pp. 3493–3514. PMLR, 2023.

Huandong Wang, Changzheng Gao, Yuchen Wu, Depeng Jin, Lina Yao, and Yong Li. Pategail: A privacy-preserving mobility trajectory generator with imitation learning. In Proceedings of the AAAI conference on artificial intelligence, volume 37, pp. 14539–14547, 2023.

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

Yusong Wang, Chuang Yang, Jiawei Wang, Xiaohang Xu, Jiayi Xu, Dongyuan Li, Chuan Xiao, and Renhe Jiang. ELLMob: Event-driven human mobility generation with self-aligned llm framework. In The Fourteenth International Conference on Learning Representations, 2026.

Takahiro Yabe, Massimiliano Luca, Kota Tsubouchi, Bruno Lepri, Marta C Gonzalez, and Esteban Moro. Enhancing human mobility research with open and standardized datasets. Nature Computational Science, 4(7):469–472, 2024a.

Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, Yoshihide Sekimoto, Kaoru Sezaki, Esteban Moro, and Alex Pentland. Yjmob100k: City-scale and longitudinal dataset of anonymized human mobility trajectories. Scientific Data, 11(1):397, 2024b.

Desheng Zhang, Juanjuan Zhao, Fan Zhang, and Tian He. Urbancps: A cyber-physical system based on multi-source big infrastructure data for heterogeneous model integration. In Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems, pp. 238–247, 2015.

Yuheng Zhang, Yuan Yuan, Jingtao Ding, Jian Yuan, and Yong Li. Noise matters: Diffusion model-based urban mobility generation with collaborative noise priors. In Proceedings of the ACM on Web Conference 2025, pp. 5352–5363, 2025.

Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James Yu. Difftraj: Generating gps trajectory with diffusion probabilistic model. Advances in Neural Information Processing Systems, 36:65168–65188, 2023.

Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Qidong Liu, Yongchao Ye, Wei Chen, Zijian Zhang, Xuetao Wei, and Yuxuan Liang. Controltraj: Controllable trajectory generation with topology-constrained diffusion model. In Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining, pp. 4676–4687, 2024.

A Detailed Methodology and Implementation

This appendix provides the complete behavioral specification, derivations, architectural details, and optimization procedure omitted from the main text for readability.

A.1 Complete Behavioral Specification

The behavioral specification contains 19 persistent user characteristics and 16 daily mobility intentions. Table 4 shows the entire feature list.

Redundant quantities - Some commonly reported mobility quantities are algebraically implied by characteristics already included above. Let NUL denote the number of unique locations and REG regularity. The corresponding number of observed visits satisfies

$$\mathrm{NoV} = \frac{\mathrm{NUL}}{1 - \mathrm{REG}}. \tag{7}$$

Likewise, because stationarity STAT determines the fraction of consecutive records that remain at the same location,

$$\mathrm{AWT} = \frac{1}{1 - \mathrm{STAT}} \tag{8}$$

for the corresponding average waiting-time quantity. Including both sides of these relationships as independently specified variables would permit internally inconsistent behavioral descriptions, so we retain the minimal specification.

Behavioral-information regimes - Daily intentions and location evidence are independent. The principal behavioral-information regimes are:

- Full: all 19 persistent characteristics and all $D \times 16$ daily intentions are supplied;
- Partial: all persistent characteristics and a subset of the daily-intention grid are supplied; and

Table 4: **Complete multi-scale behavioral specification.** Persistent characteristics describe behavior across the complete observation window, while daily intentions describe its realization on individual days.

Scale	Group	Feature	Definition / interpretation
Persistent	Profile	Intermittency	Sum of the mean consecutive exploration- and return-run lengths, capturing how long an individual tends to remain in one mobility mode before switching.
		Degree of return	Balance between consecutive returns and explorations; larger values indicate a stronger tendency toward familiar locations.
	Rhythm	Regularity	$1 - N_{\text{unique}}/N_{\text{records}}$; higher values indicate repeated use of a relatively small repertoire of locations.
		Stationarity	Fraction of consecutive observed records that remain at the same location, measuring persistence at a place over time.
		Entropy	Entropy rate of the chronological location sequence, measuring uncertainty in mobility.
		Predictability	Maximum theoretical predictability of the next location derived from trajectory entropy and the number of unique locations.
		Diversity	Normalized diversity of distinct subsequences in the location sequence, capturing variety in sequential mobility patterns.
		Dwell-time variation	Log-transformed standard deviation of consecutive dwell-run lengths, measuring variation in time spent at locations.
	Spatial	Radius of gyration	Log-transformed visit-weighted radius of gyration, measuring overall geographic spread around the mobility center.
		Unique locations	Log-transformed number of distinct locations visited across the observation window.
		k_2 ratio	Ratio between the radius of gyration of the two most visited locations and the overall radius of gyration, measuring concentration around dominant places.
		Maximum displacement	Log-transformed maximum distance between consecutive observed locations, capturing the largest single movement.
		Mean jump length	Log-transformed mean distance between consecutive observations at different locations.
		Jump-length variation	Log-transformed standard deviation of non-zero consecutive movement distances.
	Anchors	Work presence	Fraction of days containing at least one visit to an inferred work anchor; zero indicates no work anchor.
		Home–work distance	Log-transformed geographic distance between the inferred home and primary work locations.
		Home–work bearing sine	Sine component of the direction from home to the primary work location.
		Home–work bearing cosine	Cosine component of the direction from home to the primary work location.
		Number of routine places	Log-transformed number of recurrent non-home and non-work locations used by the individual.
Daily	Semantic	Home fraction	Fraction of observed slots during the day assigned to the HOME state.
		Work fraction	Fraction of observed slots during the day assigned to the WORK state.
		Exploration fraction	Fraction of observed slots during the day assigned to the EXPLORE state.
		Routine-other fraction	Fraction of observed slots spent at familiar locations other than home or work.
		Return-home indicator	Binary indicator of whether the individual returns home after leaving during the day.
		Number of places	Log-transformed number of distinct locations visited during the day.
	Temporal	Departure-time sine	Sine component of the circular encoding of the first departure time from home.
		Departure-time cosine	Cosine component of the circular encoding of the first departure time from home.
		Return-time sine	Sine component of the circular encoding of the final return time to home.
		Return-time cosine	Cosine component of the circular encoding of the final return time to home.
		Active fraction	Fraction of the day spent away from the dominant inactive or home period.
		Number of trips	Log-transformed number of trips made during the day.
	Spatial	Daily radius of gyration	Log-transformed radius of gyration computed from locations visited during that day.
		Travel distance	Log-transformed total distance accumulated between consecutive locations during the day.
		Dominant-bearing sine	Sine component of the dominant direction of movement during the day.
		Dominant-bearing cosine	Cosine component of the dominant direction of movement during the day.

- PERSISTENT-ONLY: only the 19 persistent characteristics are supplied and day-to-day behavior is generated by the model.

Separately, the recurrent-place repertoire may be fully known, partially known, or entirely absent. Location evidence therefore describes where this particular individual is known to go, rather than providing additional behavioral characteristics.

De-novo generation - For generation without an underlying real individual, we fit a Gaussian copula to the persistent-characteristic vectors of the training population and sample

$$\tilde{\mathbf{c}}^u \sim p_{\text{copula}}(\mathbf{c}^u). \tag{9}$$

The sampled characteristic vector is then passed through the persistent-only generation path.

A.2 HOME LOCATION MODEL

The Home Location Model is city-specific and trained separately from the core trajectory generator.

Residential support - Candidate homes are drawn from a mixture of the empirical training-home distribution and a spatially smoothed estimate:

$$p_{\text{home}}(c) = (1-\epsilon)\, p_{\text{emp}}(c) + \epsilon\, p_{\text{KDE}}(c;\sigma). \tag{10}$$

The empirical term preserves observed residential concentration, while the smoothed term extends support to plausible cells not occupied by a training user. The mixture weight ϵ and bandwidth σ are chosen using held-out likelihood. Each candidate cell c is represented by $\boldsymbol{\phi}(c) \in \mathbb{R}^5$ containing: (i) radial percentile relative to the city center, (ii) local residential-density percentile, (iii) local occupied-cell-density percentile, and (iv–v) the sine and cosine of its direction from the city center. Percentile normalization makes these descriptors comparable across cities.

A neural scorer

$$s_\psi\left(\mathbf{c}^u, \boldsymbol{\phi}(c)\right) \tag{11}$$

is trained with noise-contrastive ranking so that the observed home of each training individual is preferred over alternative residential candidates. At generation time, a candidate set $\mathcal{H}_u$ is sampled

from equation 10 and the final home is sampled according to

$$p_\psi(h_u = c \mid \mathbf{c}^u) = \frac{\exp s_\psi(\mathbf{c}^u, \boldsymbol{\phi}(c))}{\sum_{c' \in \mathcal{H}_u} \exp s_\psi(\mathbf{c}^u, \boldsymbol{\phi}(c'))}. \tag{12}$$

The resulting home descriptor $\mathbf{h}_u = \boldsymbol{\phi}(h_u)$ is provided independently to the Place Layout Generator and the Trajectory Decoder.

A.3 Canonical Recurrent-Place Generation

For individual u, the recurrent-place repertoire is

$$\mathcal{L}_u = \{\ell_{u,p}\}_{p=0}^{P-1}, \tag{13}$$

with

$$\ell_{u,p} = (\Delta x_{u,p}, \Delta y_{u,p}, n_{u,p}, e_{u,p}, r_{u,p}), \tag{14}$$

where $(\Delta x, \Delta y)$ is the coordinate relative to home, n is expected visitation count, e denotes whether the slot exists, and r denotes its semantic role. The canonical slot order is HOME, up to two WORK anchors, followed by remaining ROUTINE locations in descending visit count. Locations visited only once are not represented as recurrent slots and are handled by the novel-location branch.

The Place Layout Generator is an autoregressive GRU. For each slot it predicts: (i) an existence Bernoulli logit; (ii) a six-component two-dimensional Gaussian mixture for the canonical coordinate; and (iii) $\log(1 + n_p)$, the expected recurrent visitation count.

Canonical frame - Training layouts are expressed relative to home, rotated to a shared reference orientation, and normalized by spatial scale. The generator therefore learns the stochastic shape of recurrent layouts rather than having to approximate known geometric transformations. At generation time, the sampled canonical location $\hat{\mathbf{l}}_p$ is transformed into the requested frame using equation 3 of the main paper. The requested home–work bearing supplies orientation, while the requested spatial extent determines scale.

Discrete structure - Work presence directly determines whether the work anchor is instantiated. The requested number of routine places is recovered from its logarithmic representation as

$$N_{\text{routine}} = \text{round}\left[\text{expm1}\left(c_{\log N_{\text{routine}}}\right)\right]. \tag{15}$$

Anchor-preserving spatial scaling - When home, work, or supplied evidence is fixed, these anchors must not be moved simply to satisfy overall spatial scale. Let W denote total visitation weight; $(Q_A, \mathbf{a})$ the weighted second and first moments of anchored places; and $(Q_F, \mathbf{f})$ the corresponding moments of free places. Scaling only the free coordinates by s gives

$$W r_g^2(s) = Q_A + s^2 Q_F - \frac{\|\mathbf{a} + s\mathbf{f}\|_2^2}{W}. \tag{16}$$

We solve the resulting quadratic for the positive root. This allows requested global spatial extent to be approached without changing fixed anchor geometry. Because the recurrent repertoire excludes one-off exploration whereas the behavioral radius of gyration is measured over the complete trajectory, a training-derived repertoire-to-trajectory correction is applied to the requested scale.

A.4 Trajectory-Level Behavioral Targets

Several persistent characteristics are global statistics of the generated trajectory and therefore cannot be directly attached to one autoregressive step. We convert compatible characteristics into quantities that can be monitored during decoding.

Let $N^\star$ denote the desired number of observed events. Using the relationship in equation 7,

$$N^\star = \min\left(T, \frac{\text{NUL}}{1 - \text{REG}}\right). \tag{17}$$

Let $E^\star$ denote the target number of exploration events. The exploration–return profile is parameterized by intermittency μ and degree of return α. The corresponding target exploration and return run

lengths are

$$m_E = \mu \frac{\cos\alpha}{\sin\alpha + \cos\alpha}, \tag{18}$$

$$m_R = \mu \frac{\sin\alpha}{\sin\alpha + \cos\alpha}. \tag{19}$$

The implied target number of exploration/return runs is

$$K^\star = \frac{E^\star}{m_E} + \frac{N^\star - E^\star}{m_R}. \tag{20}$$

Stationarity determines the desired number of repeated consecutive observations:

$$S^\star = \mathrm{STAT}(N^\star - 1). \tag{21}$$

Together with the corresponding repertoire-size and related quantities, these targets form the compact trajectory-level representation supplied to the planner, decoder initialization, and progress tracker.

A.5 Multi-Day Planner

The Multi-Day Planner is a 4-layer, 4-head, pre-normalized Transformer encoder operating over the D day tokens. For day d, its input contains: the persistent user characteristics, the supplied daily intentions $\mathbf{c}_d^d \odot \mathbf{m}_d^d$, the mask $\mathbf{m}_d^d$, calendar/time information, and the trajectory-level behavioral targets.

It produces

$$(\mathbf{h}_1^d, \ldots, \mathbf{h}_D^d) = f_{\mathrm{plan}}\left(\mathbf{c}^u, \mathbf{C}^d \odot \mathbf{M}^d, \mathbf{M}^d, \mathbf{w}, \mathbf{b}_u\right), \tag{22}$$

where $\mathbf{b}_u$ denotes the trajectory-level targets. The planner is intentionally bidirectional. All supplied daily intentions for the complete generation window are known before the first hour is generated, so future requested days need not be hidden when forming the representation of an earlier day. An auxiliary mixture head predicts masked daily intentions as a representation-learning objective. Its samples are not inserted into the generation path: when intentions are absent, their realization remains the responsibility of the trajectory decoder.

A.6 Place Encoding and Trajectory Decoder

Each recurrent place is encoded into a fixed pointer key using an MLP:

$$\mathbf{k}_p = f_{\mathrm{place}}\left(\Delta x_p, \Delta y_p, \log(1 + n_p), e_p, r_p\right). \tag{23}$$

During training, n_p is the visitation count extracted from the target recurrent-place table, whereas at generation time this input is supplied by the Place Layout Generator's predicted count. We evaluate this train–generation asymmetry directly in Appendix G. The same encoder is used for generated recurrent places and supplied spatial evidence, so both enter the temporal decoder through the same representation. The Causal Trajectory Decoder is a 3-layer GRU spanning all $T = 240$ hours. Its hidden state is initialized from the persistent characteristics, trajectory-level targets, and home context:

$$\mathbf{H}_0 = f_{\mathrm{init}}\left(\mathbf{c}^u, \mathbf{b}_u, \mathbf{h}_u\right). \tag{24}$$

At hour t, its input is

$$\begin{aligned}\mathbf{z}_t = [&\mathbf{h}_{d(t)}^d; \mathbf{e}_t^{\mathrm{hour}}; \mathbf{e}_{d(t)}^{\mathrm{day}}; \mathbf{e}(y_{t-1}); \mathbf{q}_t^{\mathrm{prog}}; \\ &\mathbf{c}^u; \mathbf{c}_{d(t)}^d \odot \mathbf{m}_{d(t)}^d; \mathbf{m}_{d(t)}^d].\end{aligned} \tag{25}$$

Providing the masked daily intentions directly prevents all day-level information from having to pass through the planner bottleneck. The decoder emits

$$y_t \in \{0, \ldots, P-1, \mathrm{NOVEL}, \mathrm{MISSING}\}. \tag{26}$$

For recurrent place p, pointer-style scoring is

$$\ell_{t,p} = \frac{\mathbf{q}_t^\top \mathbf{k}_p}{\sqrt{d_k}} + b_{t,p}, \tag{27}$$

where $b_{t,p}$ depends on semantic role, whether the location has already been visited, cumulative visits, and recency. This preserves a fixed place identity while allowing its desirability to change as the trajectory evolves.

A.7 Behavioral Progress Tracking

After every emitted hour, the Behavioral Progress Tracker updates causal statistics including: (i) number of observations emitted; (ii) number of distinct places discovered; (iii) exploration and return counts; (iv) exploration–return run structure; (v) repeated consecutive stays; (vi) current exploration/return run length; (vii) current dwell length; and (viii) remaining time in the current day and complete trajectory. The tracker exposes not only accumulated values but normalized quantities still owed. For a generic target $B^\star$, quantity B_t already realized, and remaining horizon $T - t$, a representative progress variable is

$$q_t = \frac{B^\star - B_t}{T - t + \epsilon}. \tag{28}$$

The exploration-specific form is given in equation 4 of the main paper. For supplied daily semantic intentions, the tracker additionally computes the remaining HOME, WORK, EXPLORE, and OTHER-ROUTINE demand within the current day. A small residual-demand adjustment is added to the corresponding token logits, encouraging the day to approach its remaining semantic intention without replacing the learned transition model.

A.8 Novel-Location Generation

When the decoder emits NOVEL, an eight-component two-dimensional Gaussian mixture predicts a home-relative coordinate. A mixture is used because novel destinations conditioned on the same trajectory state are inherently multimodal; direct L_1 or L_2 regression would instead favor a central location between plausible destinations. Multiple candidates are sampled from the mixture. The behavioral specification also describes jump behavior, so candidates are evaluated according to how they reduce the remaining discrepancy in mean jump length, jump-length variation, and maximum consecutive displacement. The mixture therefore supplies plausible stochastic destinations while the remaining behavioral demand influences which candidate is realized.

A.9 Semantic State Construction

A separate semantic-state head is not learned. Instead, semantic state is derived deterministically from the emitted place and its visitation history:

$$s_t = \begin{cases} \text{HOME}, & y_t \text{ denotes home}, \\ \text{WORK}, & y_t \text{ denotes a work anchor}, \\ \text{EXPLORE}, & y_t \text{ is a first visit to a non-seeded place}, \\ \text{OTHER_ROUTINE}, & \text{otherwise}, \\ \text{MISSING}, & y_t = \text{MISSING}. \end{cases} \tag{29}$$

This ensures that the semantic history used by the Progress Tracker cannot contradict the generated spatial sequence.

A.10 City-Aware Grounding Details

The generated trajectory is represented using abstract place identities and home-relative coordinates. Grounding maps this representation to cells of the target city. Home is grounded first. If home is supplied as evidence, the observed cell is preserved; otherwise the cell selected by the Home Location Model is used. Work anchors are grounded next, in a fixed order, to available cells nearest their intended positions and are then frozen so that later destination preferences cannot overwrite the requested anchor geometry. For a routine or novel coordinate $\mathbf{q}$, let $\mathcal{N}_K(\mathbf{q})$ be the K nearest currently available city cells. equation 5 is evaluated over this set:

$$p(c \mid \mathbf{q}) = \frac{\exp\left[\lambda \log P_{\text{visit}}(c) - d(\mathbf{q}, c)/\tau\right]}{\sum_{c' \in \mathcal{N}_K(\mathbf{q})} \exp\left[\lambda \log P_{\text{visit}}(c') - d(\mathbf{q}, c')/\tau\right]}. \tag{30}$$

The neighborhood size K prevents the attractiveness prior from moving a generated destination arbitrarily far from its intended coordinate. Sampling, rather than taking the maximum-probability cell, preserves population-level destination diversity. Distinct-place semantics are maintained throughout grounding: different recurrent identities cannot collapse onto the same physical cell, and newly explored cells are not reused as independent novel destinations.

Table 5: Implementation-level architecture of *CTGTrace*.

Component	Primary input	Output	Implementation
Home Location Model	persistent characteristics, city candidates	home cell	city-specific neural scorer
Place Layout Generator	persistent characteristics, home, day summary	recurrent-place table	autoregressive GRU + 6-component GMM
Multi-Day Planner	behavioral specification	D day contexts	4-layer, 4-head Transformer
Place Encoder	recurrent-place attributes	pointer keys	MLP
Behavioral Target Module	behavioral specification	trajectory/day targets	deterministic
Progress Tracker	generated prefix	spent/remaining behavior	causal statistics
Trajectory Decoder	day context, place keys, home, progress	one location token/hour	3-layer GRU
Novel-Location Head	decoder state, spatial demand	novel coordinate	8-component 2D GMM
City Grounding	generated coordinates, city support	final cell sequence	density-aware spatial grounding

A.11 Training Objective

The Home Location Model is optimized separately using the ranking procedure in Appendix A.2. The remaining learned modules are subsequently optimized jointly. There is no user-level or day-level variational residual and no KL regularization term in the generation path.

The complete objective is

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{token}} + \mathcal{L}_{\text{exist}} + 0.15\,\mathcal{L}_{\text{novel}} + 0.15\,\mathcal{L}_{\text{layout}} \\ + 0.15\,\mathcal{L}_{\text{count}} + 0.05\,\mathcal{L}_{\text{plan}}.\end{aligned} \tag{31}$$

The terms are: (i) $\mathcal{L}_{\text{token}}$: cross-entropy over recurrent-place, NOVEL, and MISSING tokens; (ii) $\mathcal{L}_{\text{exist}}$: binary cross-entropy for recurrent-place existence; (iii) $\mathcal{L}_{\text{novel}}$: Gaussian-mixture negative log-likelihood for novel-location coordinates; (iv) $\mathcal{L}_{\text{layout}}$: Gaussian-mixture negative log-likelihood for canonical recurrent-place coordinates; (v) $\mathcal{L}_{\text{count}}$: L_1 supervision of logarithmic recurrent visitation counts; and (vi) $\mathcal{L}_{\text{plan}}$: auxiliary likelihood for masked daily intentions. Training uses AdamW with learning rate 8×10^{-4} and batch size 128. The main checkpoints are trained for 150 epochs.

A.12 Daily-Intention Masking Curriculum

A single checkpoint is trained to operate under multiple levels of daily information. For each training example, the daily-intention matrix is masked according to a mixture of patterns: (i) all daily intentions absent; (ii) subsets of complete days supplied; (iii) subsets of behavioral dimensions supplied across all days; (iv) independently scattered day–dimension entries supplied; and (v) the complete daily specification supplied. The curriculum prevents the model from associating partial information with a single fixed missingness pattern and enables the same checkpoint to operate under complete, partial, and persistent-only specifications.

A.13 Home-Context Dropout

During core-model training, residential context is removed for a random subset of training examples to reduce brittle dependence on home information. The same dropout decision is applied simultaneously to both consumers of home context—the Place Layout Generator and the Trajectory Decoder—so the spatial and temporal branches always agree on whether residential information is available.

A.14 Architecture Dimensions

The core generator contains approximately 12.6M parameters. The Home Location Model is city-specific, is trained separately, and is not included in this count.

The Multi-Day Planner has width 384. The Place Encoder produces 192-dimensional recurrent-place keys. The Causal Trajectory Decoder is a 3-layer GRU of width 512. Further dataset-specific grounding hyperparameters and sampling settings are reported with the experimental configuration.

B Datasets and Preprocessing

All main experiments use ten consecutive days with 24 hourly slots per day. Train, validation, and test partitions are user-disjoint. Table 6 summarizes the resulting population sizes, splits, observation sparsity, recurrent-place capacity, and spatial resolution used throughout the evaluation.

Table 6: Processed populations and user-disjoint splits.

Population	Users	Train	Validation	Test	Missing	Place cap.	Grid
Shanghai	58,473	40,934	8,772	8,767	.250	48	200 m
Shenzhen	12,738	8,917	1,911	1,910	.514	48	200 m
YJMob-A	68,044	47,631	10,206	10,207	.521	128	500 m
YJMob-B	18,004	12,602	2,701	2,701	.489	128	500 m
YJMob-C	16,892	11,824	2,534	2,534	.548	128	500 m

Table 7: Published generators used in the common benchmark.

Method	Venue	Family	Native generated object
MIRAGE Deng et al. (2025b)	KDD'25	neural EPR/TPP	mobility events
Marionette Deng et al. (2025a)	KDD'25	TPP + diffusion	contextual mobility events
PATEGAIL++ Ma et al. (2026)	ICLR'26	imitation learning	grid state–action sequence
DiffTraj Zhu et al. (2023)	NeurIPS'23	diffusion	GPS-coordinate trajectory
ControlTraj Zhu et al. (2024)	KDD'24	conditional diffusion	road/GPS trajectory
UPGLD Song et al. (2024)	TKDD'24	latent diffusion	discrete locations
CoDiffMob Zhang et al. (2025)	WWW'25	diffusion	continuous trajectory
DITRAS Pappalardo & Simini (2018)	DMKD'18	mechanistic	hourly location sequence
TimeGeo Jiang et al. (2016)	PNAS'16	mechanistic	temporal location sequence
EPR Song et al. (2010a)	Nat. Phys.'10	mechanistic	exploration/return sequence

As shown in Table 6, Shanghai and Shenzhen are cellular-network-derived populations. Shanghai is a non-public, anonymized telecom-derived mobility dataset collected in metropolitan Shanghai over ten days. User identities and cell-tower infrastructure are not included; locations are represented through spatial grid cells derived from the underlying network observations. Shenzhen is a publicly available anonymized cellular mobility dataset Zhang et al. (2015). Shenzhen is substantially sparser, with approximately half of its hourly slots unobserved. This matters for reconstruction-based quantities because an apparently simple behavioral statistic can be unrecoverable when the underlying trace is sparse. YJMob100K Yabe et al. (2024b) is originally provided at 30-minute resolution on a 500 m grid. YJMob-A comes from the first business-as-usual cohort. YJMob-B and YJMob-C use the same second source cohort under business-as-usual and emergency conditions, respectively. Thus A versus B changes cohort, whereas B versus C changes mobility scenario within the same dataset family. For the main protocol, YJMob is evaluated on the common hourly representation. Appendix I separately evaluates the native 48-slot half-hour representation. The semantic state vocabulary is

$$\mathcal{S} = \{\text{HOME}, \text{WORK}, \text{OTHERROUTINE}, \text{EXPLORE}, \text{MISSING}\}.$$

For each trajectory, we infer HOME as the modal cell during 02:00–05:00 and up to two WORK anchors as the modal non-home cells during 10:00–11:00 and 14:00–16:00, respectively; occurrences of these anchor cells are then assigned their corresponding semantic states. We retain WORK as a distinct state because whether and how an individual visits work directly shapes both persistent characteristics, such as work presence and home–work geometry, and daily intentions, such as work allocation and departure–return behavior. Missing observations are represented explicitly rather than forward filled. The recurrent-place capacity is $P = 48$ for Shanghai/Shenzhen and $P = 128$ for YJMob, as reported in Table 6. HOME occupies the first recurrent slot, followed by available WORK anchors and recurrent non-work locations. One-off exploration is generated through the novel-location branch rather than occupying a permanent recurrent slot.

C BENCHMARK REPRODUCTION PROTOCOL

C.1 PUBLISHED GENERATORS

The common benchmark covers ten published generators spanning neural temporal point processes, imitation learning, diffusion models, and classical mechanistic mobility processes. Table 7 lists the complete benchmark set together with each method's venue, model family, and native generated object. For every method in Table 7, each benchmark release contains $n = 1{,}000$ individuals per population. All outputs are converted to a common 10×24 city-cell sequence with an explicit missing value when the generator emits no observation. The same feature extractor, realism scorer, spatial evaluator, downstream task code, diversity analysis, and membership-inference audit are then applied to every corpus. MIRAGE, DITRAS, TimeGeo, and EPR provide the cleanest primary comparisons because their native generated objects in Table 7 already represent repeated individual places or temporal location sequences. The remaining methods require stronger representation or

Table 8: Important adaptations used by the common benchmark.

Method	Native object	Adaptation
MIRAGE	continuous-time discrete-place events	extended to ten days; event places mapped to city cells and projected onto the common hourly representation
DITRAS	hourly discrete-place sequence	city-specific tessellation, density, and mobility statistics fitted to each training population
TimeGeo	home/work/other temporal sequence	city-specific temporal/spatial parameters; output aggregated to the common hourly representation
EPR	exploration/preferential-return process	exploration destinations defined on the common city tessellation and population-specific mobility statistics fitted on training data
PATEGAIL++	grid state–action policy	common grid vocabulary and 240-step generation horizon
DiffTraj	continuous GPS trajectory	generated coordinates snapped to the common spatial grid and hourly sequence
ControlTraj	conditional road/GPS trajectory	continuous output mapped to the common city grid and hourly sequence
CoDiffMob	continuous trajectory	coordinates discretized onto the common city grid
Marionette	contextual continuous-time mobility events	horizon and category vocabulary translated to the common representation; training budget adapted to available population sizes
UPGLD	latent discrete-location sequence	location vocabulary adapted to each city and converted to the common horizon

task transfer. They are retained for field coverage, but no central conclusion relies on their degraded transferred-port behavior.

C.2 Method-Specific Adaptations

Because the native generated objects in Table 7 differ substantially, each method requires an explicit conversion into the common evaluation representation. Table 8 summarizes these adaptations and distinguishes close representation matches from transferred references.

As indicated by the final column of Table 8, MIRAGE, DITRAS, TimeGeo, and EPR require comparatively limited representation transfer. The remaining methods are therefore treated as broader field-coverage references. For PATEGAIL++, generation-quality comparisons use its zero-noise arm so that trajectory quality is not confounded with the extra distortion introduced by its privacy mechanism. Several transferred ports in Table 8 produce degenerate behavior on at least one population. Additional training-budget checks were conducted where feasible. These checks characterize the evaluated port rather than establishing a claim about the original published system, and no main-paper conclusion is based solely on such failures. All methods include repeated generation and multi seed testing.

D Complete Evaluation Protocol

D.1 Individual Behavioral Alignment

For requested factor $x_{u,j}$ and realized factor $\widehat{x}_{u,j}$,

$$\mathrm{MAE}_j = \frac{1}{N}\sum_{u=1}^{N}\frac{|x_{u,j}-\widehat{x}_{u,j}|}{\sigma_j},$$

where σ_j is the corresponding training-population standard deviation. We additionally report Pearson correlation r, regression slope, and normalized bias. A strict pass diagnostic, retained only for supplementary debugging, requires

$$r_j \geq 0.95, \qquad b_j \in [0.8, 1.25], \qquad |\mathrm{bias}_j| \leq 0.25, \qquad \mathrm{MAE}_j \leq 0.35.$$

We do not use the number of passing factors as a headline because this thresholded statistic changes materially for factors close to the boundary. The paper's 35-dimensional score is merged: the 19 persistent quantities are evaluated under persistent-only generation, while the 16 daily quantities are evaluated when those daily intentions are explicitly supplied. The complete per-population realization results are reported earlier in Table 1, while the per-factor decomposition is given in Table 10.

D.2 Population Feature Fidelity

For factor j,

$$KS_j = \sup_x \left|F_j^{\mathrm{real}}(x) - F_j^{\mathrm{syn}}(x)\right|.$$

The statistic is normalized by a matched real-versus-real sampling reference:

$$\widetilde{KS}_j = \frac{KS_j}{Q_{0.95}\left(KS_j^{\mathrm{real-real}}\right)}.$$

The benchmark compares equal-size populations, so values near one indicate discrepancy of approximately the same scale as another real sample. Normalized values from different null constructions or sample sizes are never mixed. Metric robustness beyond this KS-based view is reported in Table 13.

D.3 Independent Sequence Realism

The realism battery uses trajectory properties not summarized directly by the 35-dimensional specification: (i) unseen length-6 motifs; (ii) unseen length-8 motifs; (iii) rare length-4 motifs; (iv) rare length-6 motifs; (v) length-4 motif total variation; (vi) length-6 motif total variation; (vii) dwell-run KS; (viii) semantic-transition total variation; (ix) daily switch-rate KS; (x) circadian state-distribution deviation; and (xi) an extreme-speed-tail diagnostic. Each is normalized by matched real-versus-real variability. The extreme-speed-tail diagnostic is unbounded and can overwhelm the mean. The headline therefore averages the ten bounded diagnostics and reports the speed tail separately. The resulting cross-method comparison is summarized in Table 14. We additionally use a classifier two-sample test on the hourly sequences. Each slot is represented without cell identity using missing/stay/first-visit indicators, movement and home distances, home occupancy, and temporal features. A two-layer bidirectional GRU distinguishes $1{,}000$ generated from $1{,}000$ real trajectories under 5-fold cross-validation and three seeds; ROC-AUC near 0.5 indicates indistinguishability.

D.4 Spatial Fidelity

We evaluate complementary city-level views: cell visitation density, G-Rank, origin–destination flow, OD cosine, frequent-location overlap, and time-resolved occupancy. These are deliberately separate from individual behavioral alignment because a population can contain realistic people while placing them incorrectly over the city. Table 16 reports all six spatial views jointly.

D.5 Independent Utility

Utility follows train-on-synthetic/test-on-real (TSTR), compared with a size-matched train-real/test-real (TRTR) reference.

Next-visit prediction - Unobserved hours are removed and consecutive records at one cell are collapsed into a stay. The next destination is represented using a causal person-relative vocabulary of frequent previously visited places, other previously visited places, and a new-place class. A two-layer GRU predicts the next visit. We report Acc@1 and, in supplementary results, Acc@5 and MRR. The complete per-population Acc@1 benchmark is given in Table 17.

Zone occupancy - All visited city cells are partitioned into 256 coordinate-based spatial zones. A downstream model predicts each zone's population share using recent occupancy and temporal context. We report R^2. The full per-population comparison is reported in Table 18.

SIR application - A district-level SIR process is simulated over population co-location. Its transmission rate is calibrated on real data and then fixed. We report curve MAPE. Because real-vs-real variation is substantial, this metric is treated as an application-level stress test rather than a fine-grained ranking criterion.

D.6 Diversity and Non-Imitation

We evaluate: (i) exact ordered-trajectory copies; (ii) nearest-training location-set Jaccard similarity; (iii) within-release pairwise location-set similarity; (iv) near-duplicate rates; and (v) distinct city-cell coverage. Every quantity is interpreted against an equal-size real reference. In particular, low similarity to training data is not automatically good: a collapsed or random release can also lie far from every training person. The nearest-training comparison is reported in Table 20, while city-level coverage is reported separately in Table 21.

D.7 Release-Level Privacy

The black-box membership-inference attack asks whether a candidate real trajectory belonged to the generator's training set using only the released synthetic population. We train shadow generators with candidates alternately included and excluded, learn an attack classifier from release-level spatial,

Table 9: Behavioral realization grouped by mechanism, pooled over five populations.

Group	Factors	MAE ↓	Mean r ↑
Analytically/structurally imposed	6	.276	.864
Closed-loop targeted	10	.129	.955
Emerent/learned	19	.167	.929
no imposed structure	7	.128	.922
All	35	.175	.925

trajectory, home, and behavioral features, and report membership AUROC on the published release. DOMIAS van Breugel et al. (2023) provides the corresponding shadow-free attack for the cross-method benchmark, with AUC 0.5 indicating chance-level inference. This is an empirical attack audit and not a differential-privacy guarantee. Checkpoint-level white-box privacy is not considered here as it doesn't fall within the scope, and we will only be releasing the framework and synthesized dataset. The results are reported in Table 22.

E Further Evaluation: Individual Behavioral Realization

E.1 Realization Mechanism Taxonomy

The shipped decode path divides the behavioral dimensions into three mechanism classes: (i) **A: analytically/structurally imposed** — a closed-form value derived from the request modifies the persistent generated structure; (ii) **C: closed-loop targeted** — a target derived from the request is tracked throughout trajectory generation; (iii) **E: emergent/learned** — the request is available to the learned generator but no explicit realization rule targets the quantity. Table 9 aggregates the released-grid alignment error within these mechanism classes.

The comparison in Table 9 shows that closed-form installation does not guarantee that the same quantity is visible after trajectory realization and city grounding. In fact, the analytically/structurally imposed group has the largest mean released-grid error, whereas closed-loop targeted factors are the most accurately realized on average. This is particularly relevant for recurrent anchors: an internally placed work or routine location affects the released factor only if the emitted trace uses the location sufficiently for the external extractor to recover it. The seven most conservative emergent quantities are

$$\{\textit{return_home_flag}, \textit{dep_sin}, \textit{dep_cos}, \textit{ret_sin}, \textit{ret_cos}, \textit{day_bearing_sin}, \textit{day_bearing_cos}\}.$$

None is directly written into persistent structure or supplied with an explicit residual-demand realization rule. As summarized in Table 9, this subset still reaches MAE 0.128 SD and mean $r = 0.922$.

E.2 Factor-Level Alignment

Table 10 expands the mechanism-level summary of Table 9 to every individual behavioral dimension. This table is the most detailed view of where request realization succeeds and where substantial error remains.

Table 10 shows that the largest residual errors occur for $k2_ratio$, maximum displacement, recurrent-place reconstruction, and bearing quantities. These are also among the quantities most affected by reconstructing persistent spatial structure from the final finite-resolution released trajectory. Conversely, regularity, stationarity, distinct-location count, and several daily temporal factors exhibit very low errors and correlations close to one. Figs. 3 and 4 show the Shanghai population feature distributions of the real held-out dataset and synthetic datasets created by *CTGTrace*, EPR, MIRAGE and Marionette generators compared to real train split.

E.3 Request-Conditioned Mechanistic Baseline

Most published generators lack a paired behavioral-request interface. We therefore construct Optimized Request-EPR by tuning EPR's person-specific parameters $(\rho, \gamma, \alpha, \beta, \tau)$ from each request alone and evaluating a fresh generated trajectory; the requester's real trajectory is never exposed to the optimizer. Because EPR cannot express the full behavioral vocabulary, paired evaluation is restricted to the four compatible requested quantities. Table 11 supplements the main-text comparison with



Figure 3: **Persistent user characteristics.** Each row is one feature: real (train) in grey, and the population named by the column in blue. Panels in a row share the x-range (given under the feature name) and the density scale, capped at three times the real peak; a curve running past the top is labelled with the multiple it reaches.

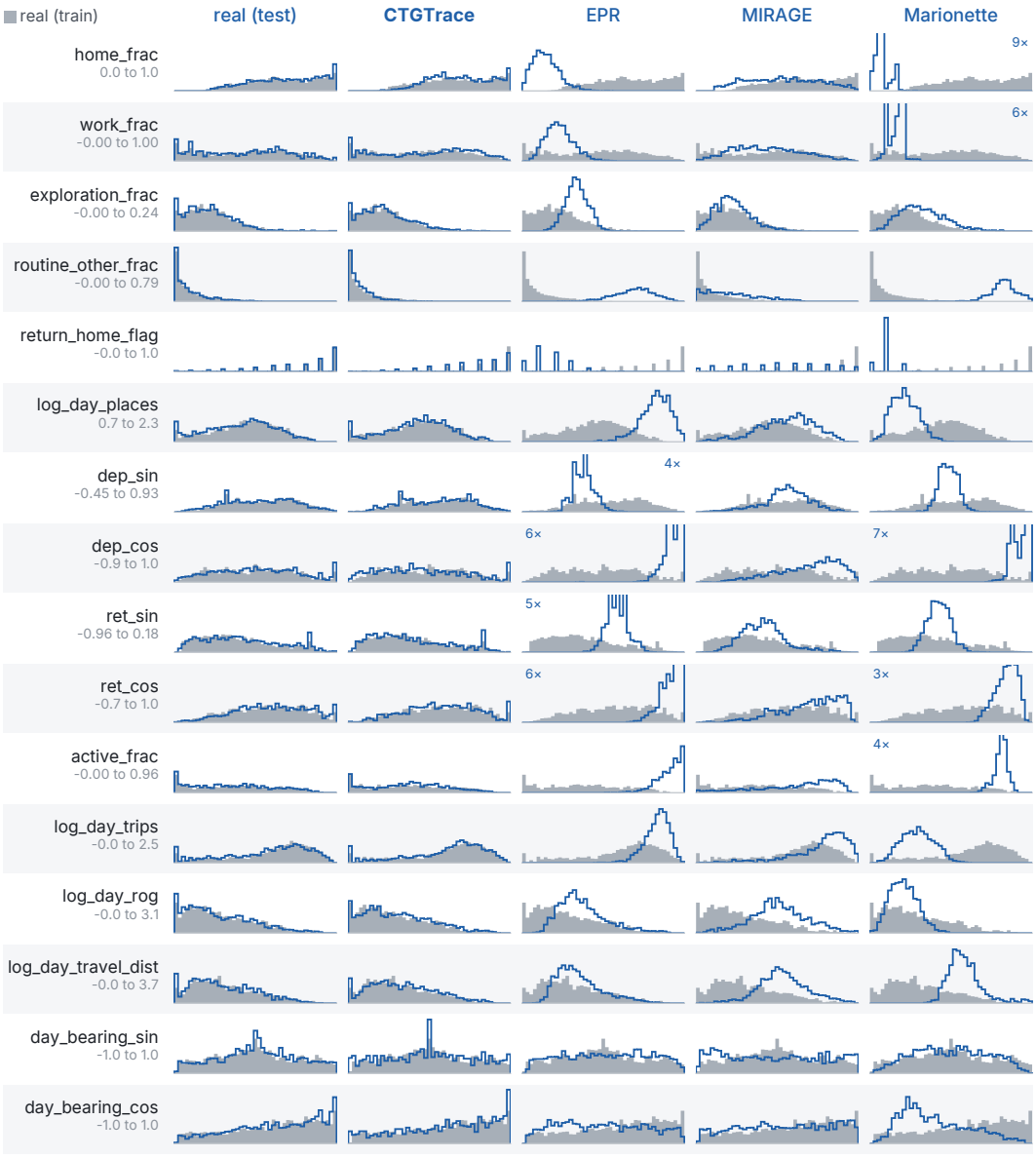

Figure 4: **Daily mobility intentions.** As above, for the 16 daily features.

Table 10: Factor-level request–realization alignment pooled across the five populations.

Factor	Type	$\text{MAE}_{\text{SD}} \downarrow$	$r \uparrow$	Slope	Bias_{SD}	Floor
intermittency	C	.075	.964	.974	-.012	.007
degree_of_return	C	.122	.983	1.031	-.043	.006
regularity	C	.044	.996	.992	-.008	.005
stationarity	C	.087	.983	.992	.006	.012
entropy	E	.169	.975	1.011	.059	.011
predictability	E	.219	.957	1.002	-.074	.012
diversity	E	.121	.958	1.105	.006	.010
log_dwell_std	E	.184	.952	.982	.030	.015
log_rog	A	.183	.974	.919	-.136	.081
log_unique	C	.051	.994	.990	.011	.006
k2_ratio	E	.506	.658	.658	.177	.125
log_mdist	E	.322	.898	.889	-.049	.118
log_jump_mean	C	.355	.953	.981	.292	.118
log_jump_std	E	.261	.931	.933	-.029	.107
work_presence	A	.357	.808	.780	.047	.044
log_home_work_km	A	.220	.850	.822	.027	.057
hw_bearing_sin	A	.260	.813	.807	-.002	.089
hw_bearing_cos	A	.261	.809	.814	.010	.081
log_n_routine	A	.376	.932	.915	-.356	.012
home_frac	C	.095	.943	.947	.035	.017
work_frac	C	.178	.884	.872	.008	.030
exploration_frac	C	.089	.963	1.015	.038	.005
routine_other_frac	C	.191	.888	.856	-.075	.022
return_home_flag	E	.132	.897	.888	.004	.015
log_day_places	C	.099	.977	.983	-.018	.008
dep_sin	E	.048	.970	.971	-.006	.009
dep_cos	E	.039	.980	.980	-.001	.010
ret_sin	E	.073	.964	.949	.021	.010
ret_cos	E	.066	.968	.958	.013	.011
active_frac	C	.060	.971	.967	-.011	.012
log_day_trips	C	.067	.979	.998	-.023	.010
log_day_rog	E	.123	.971	.905	-.033	.060
log_day_travel_dist	E	.143	.962	.899	-.016	.066
day_bearing_sin	E	.275	.827	.820	.002	.118
day_bearing_cos	E	.263	.848	.856	.002	.115

Table 11: **Requester-specific EPR: paired alignment and release quality.** Matched-4 MAE and profile agreement summarize paired request realization; the remaining columns evaluate the resulting complete synthetic release.

Method	Matched-4 MAE ↓	Profile ↑	KS×null ↓	Realism ↓	Next Acc. ↑	Occ. R^2 ↑
CTGTrace, persistent	**.150**	**.930**	**1.41**	**.99**	**.545**	**.912**
Optimized Request-EPR	.480	.753	8.28	8.11	.481	.729
Optimized Request-EPR, mask matched	0.396	0.766	4.56	4.97	.485	.901

mean profile agreement and release-level quality. Requester-specific calibration therefore improves EPR's ability to follow a request but does not recover the broader release quality of *CTGTrace*; mask matching largely repairs occupancy while leaving substantial population-fidelity and sequence-realism gaps.

E.4 Architecture-Isolating Flat Transformer Baseline

We test whether the behavioral specification can be realized by direct conditional sequence generation without the structured components of *CTGTrace*. The Flat Conditional Transformer (FCT) is a 6-layer decoder-only Transformer (d=256, 8 heads) that receives the same persistent characteristics, masked daily intentions, temporal context, and home selected by the Home Location Model, but directly predicts city cells without the recurrent-place scaffold, Multi-Day Planner, pointer mechanism, Progress Tracker, or city grounding. FCT+Prog. adds only *CTGTrace*'s causal progress representation. Both baselines use the same data splits and daily-intention masking curriculum as *CTGTrace*. As shown in Table 12, direct conditioning is substantially weaker than structured generation. FCT incurs 2.96–10.59× the persistent-request error of *CTGTrace*. Adding progress consistently reduces this error, closing 22–48% of the FCT-to-*CTGTrace* gap, but FCT+Prog. still remains 2.48–5.96× worse. Thus, progress feedback is useful, while most of the behavioral-realization advantage requires the remaining structured generation pipeline.

Table 12: Architecture-isolating behavioral realization. Persistent-19 MAE is normalized by the training-population standard deviation. "Gap" is the fraction of the FCT-to-*CTGTrace* error gap removed by adding progress. Lower MAE and ratio are better.

Population	FCT	FCT+Prog.	*CTGTrace*	FCT/Ours	Prog./Ours	Gap
Shanghai	.556	.468	**.155**	3.59×	3.02×	22%
Shenzhen	.550	.460	**.186**	2.96×	2.48×	25%
YJMob-B	2.807	1.580	**.265**	10.59×	5.96×	48%

Table 13: Mean discrepancy across thirteen conventional mobility distributions. Lower is better.

Method	Hellinger	Jensen–Shannon	Wasserstein/IQR	KS
Real sample	.069	.078	.034	.021
***CTGTrace* de-novo**	**.093**	**.107**	**.103**	**.058**
MIRAGE	.217	.249	.465	.212
PATEGAIL++	.608	.630	5.635	.582
DiffTraj	.625	.664	5.144	.642
CoDiffMob	.709	.739	6.627	.722
Marionette	.636	.676	4.294	.661
UPGLD	.672	.707	4.447	.690
ControlTraj	.701	.736	5.923	.720
DITRAS	.345	.377	.656	.293
TimeGeo	.345	.382	1.146	.338
EPR	.295	.325	.556	.264
Retrieval-19	.073	.083	.032	.023

Table 14: Independent sequence realism. The headline score excludes the unbounded extreme-speed-tail diagnostic. Lower is better.

Method	Realism-10 ↓	Speed tail ↓
Real held-out	.90	–
CTGTrace, full	.93±.03	1.3±.3
CTGTrace, persistent	.99±.02	.6±.3
CTGTrace, de-novo	1.42±.08	1.4±.4
MIRAGE	3.58±.12	26.9±2.5
DITRAS	5.60±.01	15.7±1.7
TimeGeo	6.90±.09	38.1±1.4
EPR	6.75±.02	4.6±.9

F Further Evaluation: Population level

F.1 Population-Fidelity Robustness

Feature KS acts on the 35 behavioral dimensions. We additionally evaluate thirteen conventional mobility distributions using multiple statistical distances in order to test whether the population-level finding depends on a particular discrepancy measure. Table 13 reports four representative metrics with different sensitivities and shows the same high-level relationship under all four representative distances. Hellinger and JS emphasize overlap of probability mass, Wasserstein measures how far misplaced mass must move, and KS measures the largest CDF discrepancy. Their agreement provides robustness to metric choice rather than four independent claims. The complete evaluation additionally includes total variation, Cramér–von Mises distance, energy distance, both KL directions, relative mean error, relative standard-deviation error, and relative 90th-percentile error. Retrieval-19 is included only as a calibration instrument. Because it returns real training people, its near-real distributional score demonstrates why population fidelity alone cannot establish generation or per-person alignment.

F.2 Independent Sequence Realism

Table 14 reports the independent realism battery for the three *CTGTrace* release regimes and the four primary external generators. The bounded ten-test aggregate is separated from the extreme-speed-tail diagnostic because the latter is unbounded. All three *CTGTrace* regimes remain substantially closer to the real-data reference than the primary external generators on the bounded realism battery. The remaining deviation of *CTGTrace* is concentrated primarily in semantic-transition and circadian structure rather than extreme movement tails. We further test raw sequence structure with the learned two-sample classifier in Table 15. Although *CTGTrace* remains distinguishable from real data, its AUC is substantially lower than the published-generator reference, confirming the same comparative advantage under a learned diagnostic while revealing residual sequence mismatch.

Table 15: **Learned sequence two-sample test.** ROC-AUC averaged across five populations; values near 0.5 indicate indistinguishability. Published refs. averages MIRAGE, EPR, DITRAS, and TimeGeo.

Release	**ROC-AUC** $\rightarrow .5$
Real vs. real	.487
CTGTrace, full	.566
CTGTrace, persistent	.606
CTGTrace, de-novo	.587
Published refs.	.993

Table 16: Aggregate spatial fidelity averaged across five populations.

Metric	Real	*CTGTrace* pers.	*CTGTrace* de-novo	MIRAGE	DITRAS	TimeGeo	EPR
OD-flow JSD ↓	.1724	**.2027**	.2012	.3969	.2378	.3637	.2086
OD cosine ↑	.8940	.8684	.8628	.5644	**.8691**	.8308	.8647
Per-hour density JSD ↓	.5702	.5877	.5959	.5579	**.5515**	.6543	.6622
Cell-density JSD ↓	.3055	.3080	.3075	**.2904**	.3100	.4359	.4121
G-Rank JSD ↓	–	.0084	**.0080**	.0121	.0174	.0429	.0752
Top-100 overlap ↑	–	.229	.233	.290	**.299**	.239	.167

Table 17: Next-visit Acc@1 under the common $n = 1{,}000$ benchmark.

Method	SH	SZ	YJ-A	YJ-B	YJ-C	Mean
Real	.549	.616	.521	.533	.522	.548
CTGTrace, full	.546	.609	.516	.530	.521	.544
CTGTrace, persistent	.547	.610	.519	.531	.522	.546
CTGTrace, de-novo	.539	.602	.517	.531	.520	.542
MIRAGE	.515	.586	.485	.499	.488	.515
PATEGAIL++	.313	.459	.450	.095	.452	.354
DiffTraj	.436	.421	.497	.340	.244	.388
CoDiffMob	.363	.409	.468	.290	.305	.367
Marionette	.489	.521	.214	.188	.213	.325
UPGLD	.449	.537	.200	.211	.197	.319
ControlTraj	.424	.257	.425	.179	.172	.292
DITRAS	.452	.460	.296	.291	.315	.363
TimeGeo	.377	.437	.463	.486	.472	.447
EPR	.432	.533	.425	.451	.423	.453

F.3 Population Spatial Fidelity

Individual behavioral fidelity does not ensure that a population uses the correct parts of the city. Table 16 therefore compares six complementary aggregate spatial statistics. *CTGTrace* is the strongest on OD-flow JSD, DITRAS is strongest among the generators on OD cosine and per-hour density, and MIRAGE is strongest on cell-density JSD. Conversely, *CTGTrace* closely preserves the ranked popularity curve. The spatial claim is therefore preservation of useful city structure, not uniform superiority.

F.4 Full Downstream Utility and Behavior Preservation

Next-Visit Prediction - Table 17 reports next-visit Acc@1 separately for all five populations and every generator in the common benchmark. The three *CTGTrace* regimes remain close to the real-data reference on every population. Persistent-only generation reaches mean Acc@1 0.546, compared with 0.548 for real training, while de-novo generation reaches 0.542. MIRAGE is the strongest external method at 0.515. The corresponding mean reciprocal rank is approximately 0.699 for real training, 0.695 for de-novo *CTGTrace*, and 0.675 for MIRAGE.

Zone-Occupancy Nowcasting - Table 18 provides the complementary city-level utility view by reporting R^2 for 256-zone occupancy nowcasting. This shows a more balanced comparison than the next-visit task. De-novo *CTGTrace* and MIRAGE both average approximately 0.909. Their relative ordering changes across populations, so we do not interpret the mean difference as evidence that either method uniformly dominates the other.

Synthetic Augmentation - Synthetic mobility can also complement rather than replace a small real training corpus. Table 19 evaluates the addition of 1,000 de-novo synthetic people to a 500-person real training set. This shows a positive next-visit gain on every population. The largest increase occurs on Shenzhen $(+0.045)$, while even the smallest improvement remains positive. We further target the rarest mobility profile in each population. Averaged across five populations, rare-profile

Table 18: R^2 for occupancy nowcasting over 256 spatial zones.

Method	SH	SZ	YJ-A	YJ-B	YJ-C	Mean
Real	.976	.977	.944	.887	.863	.929
CTGTrace, full	.967	.969	.919	.838	.802	.899
CTGTrace, persistent	.975	.968	.919	.852	.816	.906
CTGTrace, de-novo	.960	.953	.930	.867	.835	.909
MIRAGE	.959	.951	.919	.879	.837	.908
PATEGAIL++	.689	.690	.878	.816	.792	.773
DiffTraj	.680	.613	.769	.281	.169	.502
CoDiffMob	.460	.641	.848	.438	.396	.556
Marionette	.729	.805	.283	.875	.041	.547
UPGLD	.219	.485	.617	.415	.670	.481
ControlTraj	.617	.584	.824	.042	.101	.434
DITRAS	.648	.761	.907	.843	.810	.794
TimeGeo	.599	.752	.908	.840	.810	.782
EPR	.617	.758	.907	.837	.805	.785

Table 19: Adding 1,000 de-novo users to 500 real training users.

Population	Real only	+ Synthetic	ΔAcc@1
Shanghai	.529	.553	+.024
Shenzhen	.570	.616	+.045
YJMob-A	.507	.523	+.016
YJMob-B	.521	.536	+.015
YJMob-C	.512	.525	+.013

Table 20: Nearest-training location-set Jaccard similarity.

Population	Held-out real	*CTGTrace* de-novo	Ratio
Shanghai	.353	.303	.86
Shenzhen	.711	.563	.79
YJMob-A	.193	.176	.91
YJMob-B	.162	.151	.94
YJMob-C	.152	.141	.93

next-visit Acc@1 increases from 0.515 with 500 real users to 0.542 with unconditional synthetic augmentation and 0.553 when the 1,000 synthetic users are sampled for the target profile; adding real users of that profile gives 0.561 as an upper reference. Thus, behavioral targeting recovers most of the remaining augmentation benefit without changing the generator.

F.5 Diversity and Non-Imitation

Exact Copies and Detector Calibration - No de-novo *CTGTrace* trajectory in the common benchmark is an exact ordered copy of a training trajectory. As a positive control, Retrieval-19 deliberately republishes real training trajectories. The detector identifies 976, 843, 989, 931, and 937 exact training copies out of 1,000 outputs on Shanghai, Shenzhen, YJMob-A, YJMob-B, and YJMob-C, respectively. Thus the absence of exact copies under de-novo generation is not explained by an insensitive test.

Nearest Training Individual - Exact-copy counts do not exclude near memorization. For generated individual u,

$$J_{\max}(u) = \max_{v \in \mathcal{D}_{\text{train}}} \frac{|C_u \cap C_v|}{|C_u \cup C_v|},$$

where C_u is the set of cells visited by the person. Table 20 compares this quantity with the corresponding similarity of an independent held-out real person to the same training set.

Table 20 shows that generated people are not systematically closer to the training set than unseen real people on any population. This quantity must nevertheless be interpreted together with diversity: a collapsed or random corpus may also be far from every training person.

City Coverage - Avoiding memorization does not guarantee adequate population diversity. Table 21 therefore measures the number of distinct city cells reached by an equal-size real and de-novo population. Shanghai exhibits the clearest concentration deficit, reaching 76% of the cells covered by an equal-size real sample. YJMob-A is substantially closer at 93%, while Shenzhen slightly exceeds the reference coverage. Combined with the within-release overlap analysis, this identifies a grounding-induced fidelity–diversity trade-off rather than evidence of training-set memorization.

Table 21: Distinct city cells reached by equal-size $n = 1{,}000$ populations.

Population	Real	De-novo	Ratio
Shanghai	3,144	2,379	.76
Shenzhen	542	577	1.06
YJMob-A	12,377	11,453	.93
YJMob-B	13,284	11,171	.84
YJMob-C	12,722	11,242	.88

Table 22: Black-box training membership inference at $n = 10{,}000$; AUC 0.5 indicates chance.

Release	Mean AUC $\to$ .5	Worst population	TPR@1% FPR
Full specification	.5069	.5196	.0097
Persistent-only	.5042	.5211	.0109
De-novo	.5026	.5145	.0112

Table 23: **Expanded behavioral ablation study.** The count-input study reports MAE with ΔMAE relative to its matched shipped arm (A0); trajectory-progress ablations report MAE with ΔMAE relative to their corresponding baselines $(.119/.167/.206)$ for Shanghai/Shenzhen/YJMob-B and additionally report the change in profile agreement. Remaining retrained and decode-time interventions report ΔMAE directly. Retrained interventions use three seeds per arm.

Group	**Intervention / metric**	**Shanghai**	**Shenzhen**	**YJMob-B**
Count-input study	A0: shipped model	.134	.180	.232
	A1: remove count from pointer key	.158 (+.024)	.228 (+.048)	.292 (+.060)
	A2: predicted count in training key	.159 (+.024)	.229 (+.049)	.277 (+.045)
	A3: remove count supervision	.774 (+.640)	.866 (+.686)	.820 (+.588)
Major retrained	Remove trajectory progress: MAE (Δ)	.229 (+.1103)	.281 (+.1137)	.331 (+.1258)
	Remove trajectory progress: ΔProfile	-.0363	-.0542	-.0783
Other retrained	Remove coordinate loss	-.0010	-.0136	+.0265
	Remove coordinate-mixture objective	+.0162	+.0156	+.0273
	Remove existence supervision	+.0099	+.0576	+.0009
	Remove daily context from Place Layout Generator	–	+.0093	+.0064
	Canonical $\to$ direct place layout	–	+.0067	+.0103
	Remove direct decoder-home context	–	+.0001	+.0035
Decode-time	Remove place-layout realization	+.0114	+.0081	+.0205
	Remove work-slot existence	+.0024	+.0127	+.0005
	Remove daily-state deficit	+.0003	+.0011	+.0010
	Remove routine-count post-processing	+.0002	+.0003	+.0001
	Remove novel-jump deficit	-.0009	+.0008	+.0126
	Remove all five channels	+.0121	+.0214	+.0316

F.6 Release-Level Privacy

Table 22 reports black-box training-membership inference using DOMIAS van Breugel et al. (2023). The attacker observes only the released synthetic population and tests whether a candidate real trajectory belonged to the generator's training set. Under this attack, all three regimes remain at chance level, indicating no detectable training-membership leakage under the black-box attack.

G Expanded Ablation Study

Table 23 consolidates the final ablation suite. Retrained interventions measure the effect of removing supervision, objectives, or architectural inputs, while decode-time interventions remove deterministic channels from the shipped checkpoint. Positive ΔMAE indicates worse behavioral realization.

Training-seed variance - Across three independently trained checkpoints per population, persistent-request MAE averages 0.221 with training-seed SD 0.0025, compared with generation-draw SD 0.0034. Thus, the large retrained ablation effects reported below substantially exceed ordinary training variability.

Recurrent-place count supervision and pointer input - Table 23 separates the effect of recurrent-place count supervision from the train–generation count-input asymmetry. Removing the count from the pointer key (A1) increases MAE by 18–26%, while replacing the true training count with the Place Layout Generator's predicted count (A2) increases MAE by 18–27%. Thus, matching the training and generation inputs does not improve performance. Removing count supervision altogether (A3) is substantially more damaging, increasing MAE by 3.5–$5.8\times$. This confirms that supervised recurrent-place usage is the primary mechanism, while the count channel provides a smaller additional benefit.

Table 24: Zone-occupancy R^2 before $\rightarrow$ after final grounding.

Population	Full	Persistent	De-novo
Shanghai	.958→.968	.961→.970	.959→.959
Shenzhen	.954→.976	.959→.977	.940→.953
YJMob-A	.890→.921	.901→.931	.902→.922
YJMob-B	.814→.833	.820→.849	.829→.850
YJMob-C	.781→.819	.792→.831	.789→.817

Table 25: Requested-profile agreement as progressively more persistent characteristics are supplied.

k supplied	SH	SZ	YJ-A	YJ-B	YJ-C
0	.580	.471	.581	.597	.556
2	.660	.607	.776	.764	.781
5	.829	.747	.780	.761	.756
10	.932	.788	.784	.889	.887
19	.972	.916	.939	.940	.928

Recurrent-place usage dominates behavioral realization - Removing visitation-count supervision produces by far the largest degradation, increasing MAE by 0.560–0.637 SD. The resulting checkpoints remain numerically active, but fail to use recurrent places with the intended frequencies, showing that plausible recurrent locations alone are insufficient.

Other components refine rather than drive alignment - The coordinate-mixture objective, canonical spatial factorization, and daily layout context provide consistent but comparatively small gains. Existence supervision is particularly important on Shenzhen, whereas direct decoder-home context is negligible. The raw coordinate-loss effect is population dependent.

Decode-time corrections are secondary - Among deterministic channels, layout realization contributes most consistently, while daily-state and routine-count corrections are negligible and work-slot and novel-jump effects are population dependent. Even removing all five channels produces much smaller degradation than removing recurrent-place count supervision.

H City-Aware Grounding Intervention

The final home-density and destination-attractiveness mechanisms are evaluated on the final 256-zone occupancy task rather than the retired busiest-cell formulation. Table 24 reports the before/after occupancy result for every population and release regime. Grounding improves aggregate occupancy on nearly every population/regime combination. The largest improvements occur on the YJMob populations and on Shenzhen. Next-visit performance changes negligibly, confirming that the intervention primarily changes where individuals are grounded rather than their local temporal dynamics. The same intervention causes a small increase in behavioral MAE on some populations. It was therefore promoted for spatial fidelity and downstream population utility rather than as a behavioral-alignment mechanism. We additionally sweep the destination-attractiveness weight λ, distance scale τ, and candidate-set size K around their selected values on the validation split. Across all populations and one-at-a-time settings, persistent-request MAE varies by at most $1.039\times$, while spatial density and coverage respond more strongly. Thus, grounding hyperparameters primarily affect population placement and are not responsible for the behavioral-alignment result.

I Robustness Studies

I.1 Partial Persistent Specifications

When only part of the persistent vector is supplied, remaining values are conditionally imputed from the training population. Table 25 shows how profile agreement changes as the caller supplies progressively more persistent characteristics. This shows that useful behavioral guidance emerges before the full 19-dimensional persistent request is supplied, although the amount of information required differs substantially across populations. The experiment measures partial-request generation rather than inference of unknown characteristics, because unspecified values are imputed.

Table 26: Hourly versus native half-hour YJMob generation.

Population	Profile: 24→48	Full-spec MAE: 24→48
YJMob-A	.932→.949	.149→.275
YJMob-B	.913→.921	.173→.286
YJMob-C	.925→.915	.160→.276

Table 27: Ten-day versus fifteen-day generation.

Population	Profile: 10d→15d	Full-spec MAE: 10d→15d
Shenzhen	.934→.911	.122→.144
YJMob-A	.932→.933	.149→.148
YJMob-B	.913→.902	.173→.162
YJMob-C	.925→.928	.160→.149

Table 28: All methods are evaluated against real YJMob-C mobility. The transfer arm uses YJMob-C requests from users disjoint from YJMob-B training. Lower is better for KS, realism, JSD, profile TV, and SIR MAPE; higher is better for Acc@1 and R^2.

Setting	**Feat. KS**	**Realism**	**Cell JSD**	**OD JSD**	**Profile TV**	**Next Acc.**	**Occ. R^2**	**SIR MAPE**
Real YJMob-C reference	.840	.713	.3859	.1357	.0304	.5247	.8678	24.2%
YJMob-C model	1.560	1.018	.4513	.1764	.0324	.5235	.8430	32.1%
YJMob-B → YJMob-C	1.518	1.068	.4513	.1750	**.0106**	.5228	.8509	38.5%

I.2 Temporal Resolution

YJMob is natively observed at 30-minute resolution. Table 26 compares the default 24-slot hourly representation with its native 48-slot representation. This shows that the finer temporal grid provides no consistent profile improvement and substantially increases fine-grained behavioral error on all three populations. This supports the hourly representation used by the main experiments.

I.3 Observation Window

Table 27 tests whether the main conclusions depend on the ten-day observation horizon by extending generation to fifteen days. The longer observation window produces only modest and inconsistent changes. YJMob-A/B/C change little in behavioral error, while the sparse Shenzhen population degrades somewhat. The main conclusions therefore do not depend on a particularly favorable ten-day horizon.

I.4 Scenario Transfer

We evaluate whether a model trained on YJMob-B business-as-usual mobility can realize YJMob-C emergency-period requests without retraining. Source and target users are disjoint for this evaluation. Table 28 shows that the transferred model remains close to the matched YJMob-C model across behavioral, sequence, spatial, and downstream measures. In particular, normalized feature KS is 1.518 versus 1.560 in-domain, while the generated profile distribution is within TV 0.011 of real YJMob-C.