Highlights

PopTraj: Hierarchical Multi-Source Generation of Household-Coordinated Weekly Activity–Location Trajectories with Urban Populations Synthesis

Haoxuan Ma, Yifan Liu, Bo Yang, Yiwen Zhang, Chris Stanford, Eli Fox, Xishun Liao, Jiaqi Ma

- A hierarchical framework fuses census, survey, geographic, and mobility data.
- A language-model teacher supplies joint schedules; distillation enables scale.
- Household scheduling cuts coordination error by more than 16-fold.
- Home-anchored trajectory generation lowers mean divergence from 0.0412 to 0.0252.
- Quantized persona latents halve demographic divergence while preserving realism.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

PopTraj: Hierarchical Multi-Source Generation of Household-Coordinated Weekly Activity–Location Trajectories with Urban Populations Synthesis

Haoxuan Ma[a,1], Yifan Liu[a,*,1], Bo Yang[a], Yiwen Zhang[b], Chris Stanford[c], Eli Fox[c], Xishun Liao[d] and Jiaqi Ma[a,*]

[a]Department of Civil and Environmental Engineering, University of California, Los Angeles, Los Angeles, CA, USA
[b]Center for Data Science, New York University, New York, NY, USA
[c]Novateur Research Solutions, Ashburn, VA, USA
[d]Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL, USA

ARTICLE INFO

Keywords:
Human mobility
Synthetic population
Trajectory generation
Urban AI
Multi-source data fusion

ABSTRACT

Synthetic human mobility trajectories support urban planning, transportation simulation, and epidemic analysis by representing when, where, and how people conduct activities. However, existing approaches often model individuals independently or over single days, limiting joint representation of overnight activity, demographic heterogeneity, household interactions, and multiday regularities. The required evidence is fragmented across census, survey, geographic, and anonymous GPS data that cannot be linked at the individual level. We present PopTraj, a hierarchical multi-source generation framework for producing household-coordinated weekly activity–location trajectories for synthetic urban populations. PopTraj organizes generation into three source-aligned levels. A sample-free synthesizer reconstructs a household population from aggregate census constraints. A survey-calibrated language model produces a coordinated-schedule teacher corpus, which is distilled into a Hybrid 3D conditional variational autoencoder for scalable generation. Finally, HEARTH completes each dictated week as home-anchored excursions, using a vector-quantized persona latent to connect GPS-learned behavioral archetypes with a survey-derived demographic prior. In a metropolitan-scale Tokyo deployment, PopTraj generates 395K synthetic person-weeks. HEARTH achieves a mean divergence of 0.0252 across six realism metrics, compared with 0.0412 for the strongest of nine baselines. The deployed schedule pipeline reduces household coordination error by more than 16-fold relative to six learned baselines, while the persona prior halves demographic divergence with little change in aggregate realism. Source-removal, architecture, and error-propagation analyses show that the persona prior supplies demographic separation and the household schedule supplies cross-level coordination. PopTraj therefore provides a scalable route from unlinked population statistics and anonymous mobility data to simulation-ready human mobility trajectories.

1. Introduction

Modern cities generate large volumes of data describing human activity, infrastructure use, environmental conditions, and transportation systems. Urban computing uses these heterogeneous signals to understand how cities function and to support planning, forecasting, and policy evaluation. Among these signals, human mobility data are especially important because they connect residents to the places where urban life occurs. Fine-grained activity–location trajectories reveal not only where people travel, but also when they work, study, shop, receive care, and participate in household and community life. Such trajectories underpin applications ranging from transportation planning and epidemic forecasting to infrastructure-stress testing and public-policy simulation [1, 2, 3].

Yet the mobility data required by these applications are difficult to collect at metropolitan scale. Detailed individual traces are costly, privacy-sensitive, and rarely accompanied by the demographic and household information needed to interpret behavior. Synthetic trajectory generation seeks to overcome this scarcity by producing realistic mobility records without requiring an observed trace for every simulated resident. The challenge is not simply to generate plausible location sequences for isolated individuals. A useful synthetic population must preserve demographic variation, coordination among household members, overnight activity, and recurring behavior across the seven-day cycle while remaining scalable to millions of residents. Constructing such a population is the problem studied in this paper.

No single data source satisfies these requirements jointly. Census tabulations report the demographic composition of a region but not its residents' time use; time-use surveys record activity participation by demographic group, but only for limited samples and through retrospective diaries rather than continuous traces [4]; and the anonymous GPS corpus used here captures detailed movement but contains no demographic labels. The signals the three properties require, demographic structure, household linkage, and zone-resolved movement over multiple days, reside in separate datasets that cannot be linked at the individual level, and a model trained on any single source inherits the corresponding blind spot. The complete weekly life a simulation needs is recorded, as

*Corresponding authors.
bmmliu@ucla.edu (Y. Liu); jiaqima@ucla.edu (J. Ma)
[1]H. Ma and Y. Liu contributed equally.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

one object, nowhere. Scale compounds the difficulty: with millions of residents to generate, every model must survive the per-resident inference loop.

Each existing line solves the part of this contract its own source can teach. Activity-based models generate demographically grounded schedules and activity programs from survey data and behavioral or agent-based mechanisms [5, 2, 1]. These systems generally require substantial calibration but do not jointly target household-coordinated, census-aligned, zone-resolved weekly synthesis for previously unobserved residents. Deep mobility generators include sequence models [6], adversarial generation and imitation [7, 8], diffusion models [9], LLM agents [10], and mobility foundation models [11]. These approaches differ substantially in task and resolution, but they do not jointly provide explicit household coordination, census-level demographic alignment, and complete weekly activity–location generation for previously unobserved residents. A third line of work completes the unobserved portions of an *observed* trajectory [12, 13, 14], but requires each individual's own trace and therefore cannot serve synthetic residents who have never been observed.

The obstacle is structural: neither the required behavioral dependencies nor the evidence sources align at the individual level. POPTRAJ therefore treats census-to-trajectory synthesis as hierarchical multi-source generation, factorized through a latent household schedule and trained level by level on the source that observes each target. Three commitments guide the design. Source-aligned factorization limits each source to observed properties. Behavior-aligned representations separate shared events from individual routines, read weeks on a grid, and generate home-anchored excursions. Deployment-aligned control handles household copresence and demographic alignment explicitly. These commitments address three challenges: synthesizing a census-faithful population (C1), coordinating household schedules (C2), and completing demographic-conditional trajectories for unseen residents at metropolitan scale (C3).

Prior systems target related but distinct tasks, so we evaluate on two complementary planes. Cross-task systems are positioned on capability axes in Table 1; for Level III we build a nine-baseline comparison on the exact deployment task, adapting published completion models where needed, while Levels I and II are validated against census and national time-use references. In summary, our key contributions are:

- We formulate population-scale trajectory synthesis from mutually unlinkable data as hierarchical multi-source generation, an evidence-routing form of multilevel fusion, and propose POPTRAJ, to our knowledge the first learned framework to generate demographically grounded, household-coordinated weekly activity–location trajectories for synthetic residents with no observed mobility history.
- We develop a source-aligned three-level architecture that connects a sample-free population synthesizer, a survey-calibrated household schedule generator, and HEARTH (**H**ome-anchored **E**pisodic **A**ctivity-location **R**econstruction **T**ransformer for **H**ouseholds), a behavior-aligned trajectory completion model. This factorization allows each data source to supervise the properties it observes while carrying demographic and household constraints across levels.
- Extensive experiments on a metropolitan-scale Tokyo deployment show that POPTRAJ preserves census population structure, household coordination, demographic differences, and GPS-derived mobility distributions. HEARTH reaches 0.0252 mean divergence versus 0.0412 for the strongest of nine trajectory baselines; the deployed schedule pipeline reduces coordination error by more than 16× relative to six learned baselines; and the persona prior halves demographic divergence with little change in aggregate realism. The framework generates hundreds of thousands of weekly trajectories in hours without a language model at inference.

2. Related Work

Three lines of work border this paper: multi-source urban mobility and population synthesis, household behavior generation, and mobility generation and completion.

2.1. Multi-source urban mobility and population synthesis

Urban computing has embraced multisource fusion as the route to operational intelligence, combining sensing, mobility, infrastructure, and environmental streams for monitoring and forecasting across smart-city applications [15]. In mobility forecasting specifically, heterogeneous-source architectures improve large-scale traffic prediction by embedding diverse urban data into shared spatiotemporal representations [16], and mobile-phone origin–destination matrices sharpen synthetic travel demand built from surveys [17]. Closest to our population layer, Vo et al. fuse household travel surveys with passively collected cellular data through cluster-based optimization, generating synthetic populations whose sociodemographics and home–work locations achieve high spatial heterogeneity [18]. These systems integrate unlinked sources for population attributes or aggregate flows. POPTRAJ extends the fused object itself: beyond home–work attributes, it generates coordinated multi-day activity–location trajectories for every member of every synthetic household, which requires the additional household and episode levels this paper formalizes.

2.2. Household activity and relational behavior generation

Activity-based travel demand systems generate daily schedules through econometric choice hierarchies, rule-based episode insertion, or co-evolutionary scheduling [5, 2, 1]. Synthetic-population resources extend these

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

systems to larger populations: SySMo provides 10.2M Swedish agents with household structure [19], patterns-of-life simulators operate at tens-of-millions scale [3], and sample-free methods such as GenSynthPop reconstruct attribute dependencies from aggregate tables [20]. Household travel has long been modeled as a joint decision problem [21], while recent deep generators such as ActVAE generate person-level schedules conditioned on individual and household attributes [22]. Existing systems do not generally combine explicit household coordination, survey-targeted demographic alignment, and weekly zone-resolved trajectory generation in one learned pipeline. LLM agents offer a complementary source of semantic priors about individual activity patterns and motivations [10]. PopTraj uses that knowledge only to construct a survey- and GPS-calibrated teacher corpus, which is then distilled into a student model that requires no language model during metropolitan-scale inference.

2.3. Human mobility generation and completion

Neural mobility generators include RNN sequence models [6], adversarial generation and imitation [7, 8], temporal point processes [23], diffusion models [9], and mechanistic–statistical hybrids such as TimeGeo [24]. MIRAGE generates event time, activity category, and location by combining an intensity-free temporal point process, a neural exploration-and-preferential-return decoder, and a user-level VAE [25]. Traveller instead uses autoregressive temporal planning to guide diffusion-based spatial generation [26]. Our earlier foundation model fused trajectory, survey, and infrastructure data for cross-domain mobility modeling [11]; PopTraj extends that line to household-coordinated weekly trajectories. A separate literature reconstructs missing portions of observed trajectories. DeepTimeGeo infills sparse location sequences [12]; BERT4Traj uses anchor and demographic context to recover CDR/GPS gaps [13]; and TrajGPT jointly models region, arrival time, and duration during controlled visit infilling [27]. VSNIT inserts missing activities [14], attention-based models predict or recover locations from mobility histories [28, 29], and semi-supervised methods transfer activity chains across datasets [30]. These systems either require the target person's observed trace or generate without household-semantic constraints. HEARTH instead completes a partial weekly household schedule for a synthetic resident who has never been observed. It preserves dictated events, predicts activities and zones jointly, represents home as an ambient state, conditions the week through a demographic persona prior, and operates inside an end-to-end census-to-trajectory pipeline. Section 7 evaluates this distinction directly using MIRAGE and task-adapted completion baselines.

3. Problem Formulation

We formulate PopTraj as a conditional generative model that produces a complete weekly activity–location trajectory for every member of a synthetic household population. Let $\mathbf{m}$ collect the regional aggregate constraints supplied to the population synthesizer, and let $\mathcal{P}$ denote the resulting synthetic population. Each person $n \in \mathcal{P}$ belongs to a household $h(n)$ with demographic vector $\mathbf{d}_{h(n)} \in \mathbb{R}^D$, which records household composition, member roles, sex, age, employment, work shift, income, vehicle ownership, and religious affiliation.

Let $\mathcal{S}_h$ denote the recurring weekly schedule of household h, including member-specific mandatory activities and coordinated shared events. Let $\boldsymbol{\eta}_n$ collect the Level-I home, Work, and School anchor zones available for person n. The final trajectory of person n is a pair $(\mathbf{a}_n, \mathbf{z}_n)$ of activity and zone labels on a shared 15-minute grid. The complete generation follows three levels:

$$\begin{aligned} \mathcal{P} &\sim p_\alpha(\mathcal{P} \mid \mathbf{m}), \\ \mathcal{S}_h &\sim p_{\phi,\theta}\big(\mathcal{S}_h \mid \mathbf{d}_h\big), \\ p(E_n \mid \mathcal{S}_{h(n)}, \mathbf{d}_{h(n)}, \boldsymbol{\eta}_n) &= \sum_{z_n^{\mathrm{p}}} \pi\big(z_n^{\mathrm{p}} \mid \mathbf{d}_{h(n)}\big) \\ &\quad \times p_\psi\big(E_n \mid \mathbf{x}_n^{\mathrm{dict}}(\mathcal{S}_{h(n)}, \boldsymbol{\eta}_n), z_n^{\mathrm{p}}\big), \end{aligned} \tag{1}$$

where z_n^{p} is a discrete persona code and E_n is the excursion sequence defined below. The three lines of Eq. 1 correspond to the three model levels. Level I generates the population and assigns home and mandatory activity anchors. Level II maps each household record to a coordinated weekly schedule. Level III combines that schedule with the Level-I anchor tuple $\boldsymbol{\eta}_n$ to form the dictated input $\mathbf{x}_n^{\mathrm{dict}}$ and generates the remaining activities and locations. The levels share the demographic vector, activity vocabulary, zone partition, and weekly grid, but their learned components share no model parameters. Time-use surveys, the teacher corpus, GPS observations, and geographic data provide training evidence or group-level targets for the corresponding components; they are not individual-level inputs required at generation time.

Each person is represented by $T = 672$ time bins, with 96 bins per day for seven days. Two additional positions hold the home zone and one primary mandatory recurring activity zone, giving total sequence length $L = 674$. Each position t contains an activity token a_t, a zone token z_t, a time-of-day index $\tau_t \in \{0, \dots, 95\}$, a day-of-week index $\delta_t \in \{0, \dots, 6\}$, and a role indicator r_t. The activity vocabulary contains 11 classes (Home, Work, School, ChildCare, EatOut, Errands, Recreation, Visit, Healthcare, Religious, and DropOff) together with `[PAD]`, `[CLS]`, `[MASK]`, and `[UNK]`. The zone vocabulary contains the $|\mathcal{Z}|$ deployment zones, `[UNK_LOC]`, and `[MASK_LOC]`. Level II produces only activity schedules: recurring Work and School episodes and household-coordinated events with their times and participant sets. Level III combines these schedules with the Level-I anchor zones as input. Work and School use the fixed Level-I activity–location anchors. For each coordinated event, HEARTH infers one location from the household head's trajectory context, and framework post-processing assigns that location to the event's other participants. Level III completes the remaining unscheduled

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

Table 1
Capability positioning of representative system families against the PopTraj contract (✓ supported; ∼ partial or restricted; × not supported). Cross-task systems are positioned on capability axes; systems adaptable to our completion task additionally enter the numeric comparison of Section 7.

Family	Systems	Unlinked sources	Aggr. census	Survey behavior	Passive mobility	Geo. sem.	Househ. relations	Multilevel latent fus.	Weekly traj.
Traj. completion	DeepTimeGeo [12], BERT4Traj [13], TrajGPT [27]	×	×	×	✓	∼	×	×	∼
Deep generative	DiffTraj [9], ActSTD [23], MIRAGE [25], Traveller [26]	×	×	×	✓	∼	×	×	×
Mechanistic	TimeGeo [24]	×	×	×	✓	∼	×	×	∼
LLM-agent	LLMob [10]	×	×	∼	∼	✓	×	×	×
Synthetic population	SySMo [19], SimPOL [3]	∼	✓	✓	∼	∼	∼	×	∼
Survey+passive synthesis	Vo et al. [18]	✓	✓	✓	✓	×	∼	×	×
PopTraj (ours)	this paper	✓	✓	✓	✓	✓	✓	✓	✓

behavior, after which rasterization fills all unoccupied bins as Home. These dictated entries cover roughly one third of the week and remain fixed in the final trajectory. Reserved activity and location tokens serve structural input roles only and never appear in completed trajectories. Supplementary Section SI summarizes this division of activity-representation responsibilities between Levels II and III in tabular form.

The free portion of the week is represented as an excursion sequence $E_n = (e_1, \dots, e_{K_n})$. Each excursion $e_k = (w_k, a_k, u_k, z_k)$ records the preceding home duration $w_k \geq 0$, its activity a_k, the excursion duration u_k, and its zone z_k. Setting $w_k = 0$ chains consecutive excursions without an intervening return home. HEARTH estimates

$$p_\psi\left(E_n \mid \mathbf{x}_n^{\text{dict}}, z_n^{\text{p}}\right) = \prod_{k=1}^{K_n} p_\psi\left(e_k \mid e_{<k}, \mathbf{x}_n^{\text{dict}}, z_n^{\text{p}}\right). \quad (2)$$

A deterministic rasterization maps the generated excursions and dictated entries back to the weekly grid. Excursions occupy their generated intervals, supplied fields are restored unchanged, and Home fills all remaining bins. During training, structured masking converts GPS-derived weeks into completion examples. At inference, HEARTH receives the Level II schedule and generates only its missing content.

4. Methodology

4.1. Framework overview and evidence routing

Figure 1 summarizes the three levels. Level I (Section 4.2) converts the aggregate constraints $\mathbf{m}$ into a synthetic household population $\mathcal{P}$. Level II (Sections 4.3–4.4) generates a coordinated weekly schedule S_h for each household record $\mathbf{d}_h$. Level III (Sections 4.5–4.6) conditions on that schedule to complete the weekly trajectories. The levels are trained separately and share no parameters. The synthetic population and household schedules carry the required context from one level to the next. Table 3 in Section 5 lists the unit, resolution, role, and reliability assumption of each data source.

4.2. Level I: Census-constrained population synthesis

Level I converts aggregate census tables into household records. Its output is the demographic vector $\mathbf{d}_h$ used by the schedule generator, the persona prior, and the demographic-steering control in Section 4.6. The target marginals cover household size (1–6 members), age in 5-year bands, sex, employment status, work shift, income percentile, vehicle ownership, and religious affiliation. Households are built through conditional sampling. The sampler draws household size, assigns a role to each member, and then draws sex, age, employment, shift, income, religion, and vehicle ownership from distributions conditioned on the attributes already assigned. Compared with earlier sequential synthesizers such as GenSynthPop [20], our method adds a detailed household role taxonomy, head-relative age distributions based on empirical family gaps, and a final correction to match the published population pyramid. These choices preserve the target marginals while recovering relationships such as spousal age gaps and dependence among age, role, and employment (Section 7.1).

The resulting record is serialized into a fixed-length demographic vector $\mathbf{d}_h \in \mathbb{R}^{122}$ with the layout

$$\mathbf{d}_h = \Big[\underbrace{\mathbf{u}_h}_{\text{household block (8 dims)}} ; \underbrace{\mathbf{v}_{h,1}, \dots, \mathbf{v}_{h,6}}_{\text{6 member blocks (19 dims each)}} \Big], \quad (3)$$

where $\mathbf{u}_h$ contains household size, income percentile, vehicle count, and a five-way religion one-hot vector. Each $\mathbf{v}_{h,m}$ contains a seven-way role one-hot vector, sex, normalized age age/100, a five-way employment one-hot vector, and a five-way shift one-hot vector over {day, evening, overnight, none, unknown}. The last two shift categories represent non-employed members and members whose shift is unrecorded, respectively. Absent members ($m > M_h$) are zero-padded. This fixed serialization lets the CVAE encoder and per-member loss heads access member-specific attributes by slice without parsing variable-length records.

Each household receives one home zone drawn from the empirical home-zone distribution. Work and school zones are then assigned to individual members with a gravity

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

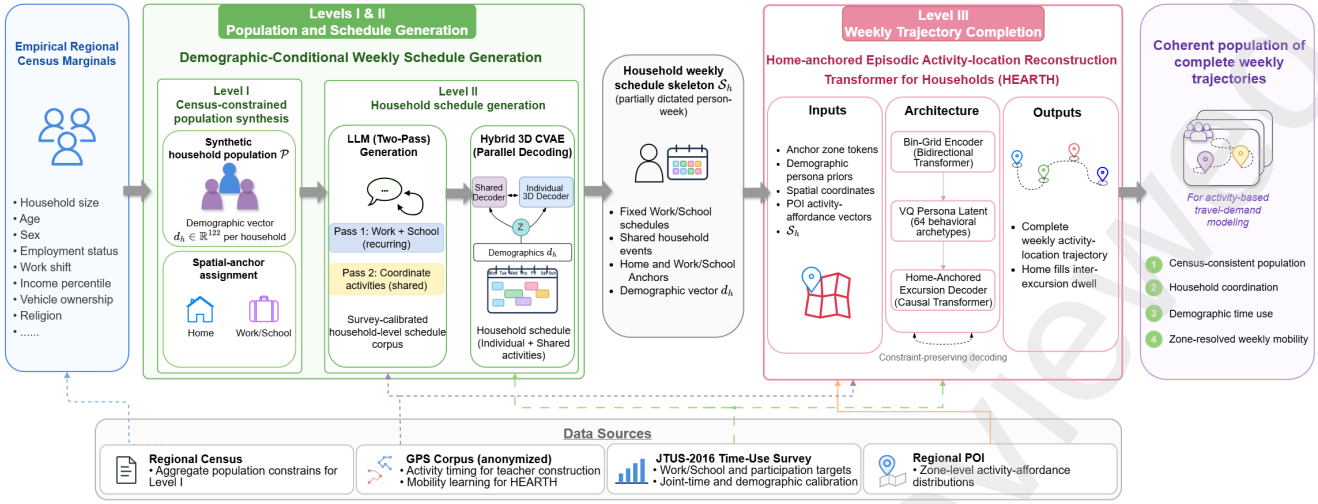

Figure 1: PopTraj as a hierarchical fusion architecture. Left: five evidence sources with their native entity levels and resolutions. Center: three levels connected by latent alignment variables, synthetic households and household weekly schedule skeletons, with arrows labeled by the evidence each level consumes. Right: the coherent synthetic object, a census-consistent, household-coordinated population with zone-resolved weekly trajectories.

model [31]. For home zone o and mandatory activity a,

$$P(d \mid o, a) \propto A_d^{(a)} \cdot e^{-\beta\, \mathrm{dist}_{od}} \cdot \mathbf{1}[d \in C_o^{(a)}], \tag{4}$$

where $A_d^{(a)}$ is zone attractiveness estimated from reference activity counts, dist_{od} is the distance between zones, and $\beta = 0.1\,\mathrm{km}^{-1}$. The candidate set $C_o^{(a)}$ restricts destinations to a band around the member's target commute distance, sampled from the empirical home-to-mandatory distance distribution. This procedure assigns realistic spatial anchors without requiring linked census and mobility records.

4.3. Level II-A: Survey- and mobility-grounded semantic teacher

The schedule generator in Section 4.4 is trained on weekly activity chains produced by a retrieval-augmented, two-pass large-language-model (LLM) pipeline. For each synthetic household, the prompt combines its demographic record $\mathbf{d}_h$ with targets retrieved from a national time-use survey (TUS) and a held-out subset of the GPS corpus. The survey provides weekly Work and School hours and daily participation rates for the matching demographic groups. The GPS data provide activity-specific time-of-day participation curves. These numerical targets are inserted as in-context constraints, while the system message fixes the activity vocabulary and output JSON schema.

Generation uses two passes. The first creates recurring Work and School episodes, allowing at most one continuous episode of each type per member per day. Survey hour and participation targets, GPS timing curves, and the work shift in $\mathbf{d}_h$ guide their placement. The second pass adds coordinated activities without overlapping the fixed Work and School schedule. Its prompt also includes survey estimates of joint time for couples and parent and child pairs, which guide the participant sets of shared activities. The two passes populate the individual and shared streams used by the CVAE (Section 4.4). Supplementary Section SII details how the JTUS day-type tables are converted into weekly recurring and dyad budgets, how the survey companion-time buckets map to the Level-II activity vocabulary, and how survey co-presence targets are translated into conformance bands for out-of-home schedules.

Each pass is checked against deterministic rules and regenerated up to six times if it fails. For recurring activities, the checks enforce employment-consistent weekly hours and prevent same-day overlaps. For coordinated activities, they enforce the activity vocabulary, prevent collisions with recurring episodes, and keep each pair's joint time within a permitted range around its survey target. After the two passes are combined, a deterministic cleanup removes remaining conflicts and restores weekend Work and School participation to GPS-calibrated rates. These steps address excess mandatory activity for mid-aged students and elderly nonworkers, as well as under-generation on weekends. The final corpus tracks survey weekly hours within a few hours across demographic groups and maintains population activity shares consistent with the survey and GPS references. The corpus composition, prompt configuration, regeneration settings, representation coverage, and conformance-failure breakdown are reported in Supplementary Section SIII.

4.4. Level II-B: Household schedule generation and distillation

The schedule generator models the Level-II factor $p_{\phi,\theta}(S_h \mid \mathbf{d}_h)$ in Eq. 1 with a household-conditioned Conditional Variational Autoencoder. Each household week has a shared stream for coordinated events and an individual stream for each member's routine. A bidirectional gated cross-fusion layer connects the two streams before decoding. Shared events can therefore respond to individual routines, and individual schedules can account for planned

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

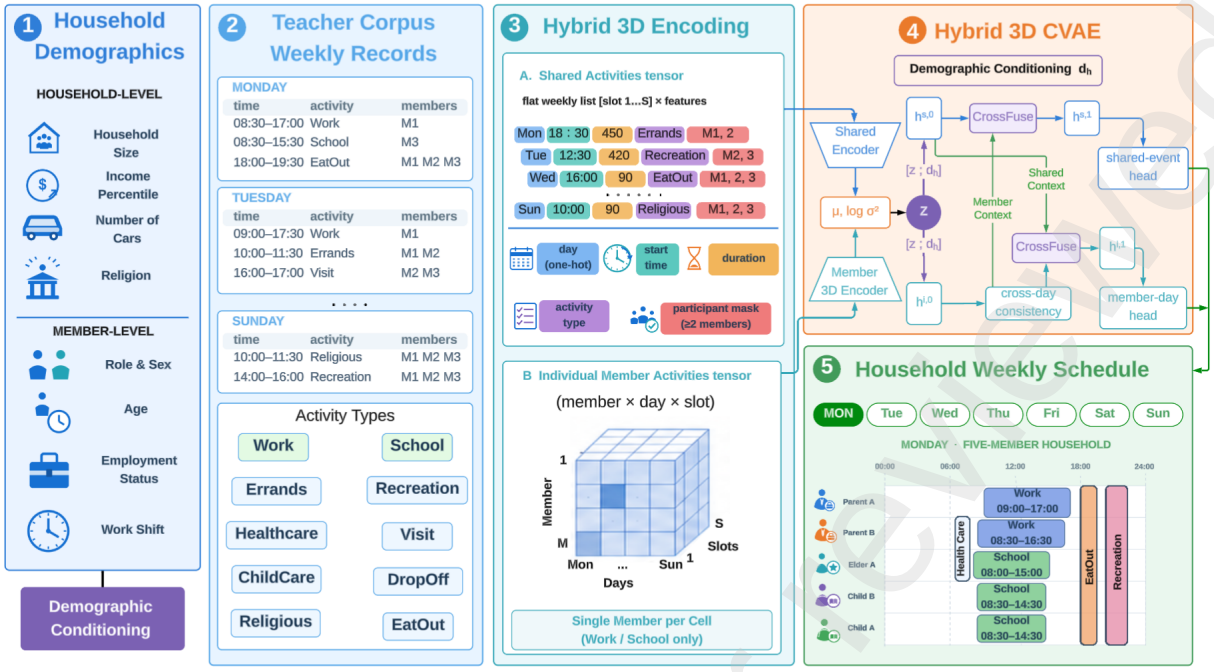

Figure 2: The Hybrid 3D Conditional VAE, the Level-II schedule generator. (1) Per-household demographic vector $\mathbf{d}_h$. (2) Weekly recurring activity records from the teacher corpus. (3) Hybrid 3D encoding: a fixed-length list of shared activities plus a [member × day × slot] tensor of individual activities. (4) Demographic-conditioned encoder, latent $\mathbf{z}$, and twin decoders that exchange information through bidirectional gated cross-fusion. (5) Decoded shared household events and member-day schedules, forming the household weekly skeleton S_h that Level III consumes.

household events. The encoder $q_\phi(\mathbf{z} \mid S_h, \mathbf{d}_h)$ produces a diagonal-Gaussian posterior with an isotropic Gaussian prior. The decoder $p_\theta(S_h \mid \mathbf{z}, \mathbf{d}_h)$ reconstructs both streams in one forward pass. Once trained, this student model replaces the LLM during large-scale generation. Figure 2 summarizes the model.

Schedule representation. Each schedule contains a shared stream and an individual stream. The shared stream stores up to $S_{\max} = 35$ coordinated activities,

$$\mathbf{s}_j = \Big[\underbrace{\mathbf{o}_j^{\text{day}}}_{7} ; \underbrace{\sin\theta_j^s, \cos\theta_j^s}_{2} ; \underbrace{\tilde{\ell}_j^s}_{1} ; \underbrace{\mathbf{o}_j^{\text{act}}}_{11} ; \underbrace{\mathbf{m}_j}_{6} \Big] \in \mathbb{R}^{27}, \quad j = 1, \dots, S_{\max}, \tag{5}$$

Each row records the day, cyclic start time, log-normalized duration, activity, and participating household members. The 11-way Level-II activity vector contains the ten non-Home schedule activities plus one padding class; Home remains implicit and is not a Level-II event. The individual stream is a tensor of shape $[M_{\max} \times D \times A] = [6 \times 7 \times 5]$, with each cell represented as

$$\mathbf{x}_{m,d}^{(k)} = \Big[\underbrace{t_{m,d}^{(k)}}_{2} ; \underbrace{\tilde{\ell}_{m,d}^{(k)}}_{1} ; \underbrace{\mathbf{o}_{m,d}^{\text{act},(k)}}_{11} ; \underbrace{\pi_{m,d}^{(k)}}_{1} \Big] \in \mathbb{R}^{15}, \tag{6}$$

where $\pi_{m,d}^{(k)}$ indicates whether the slot is active. Time is encoded in two forms:

$$t_{m,d}^{(k)} = \begin{cases} (\sin\theta_{m,d}^{(0)}, \cos\theta_{m,d}^{(0)}) & k = 0, \\ (\text{gap}_{m,d}^{(k)}/12, 0) & k \geq 1, \end{cases} \tag{7}$$

The first slot stores an absolute start time, while later slots store the non-negative gap from the preceding activity. Enforcing $\hat{\text{gap}}_{m,d}^{(k)} \geq 0$ during decoding prevents activities within the same day from overlapping.

Encoder, decoders, and cross-fusion. The encoder flattens both streams, processes them with stream-specific MLPs conditioned on $\mathbf{d}_h$, and combines them into a diagonal-Gaussian posterior over $\mathbf{z} \in \mathbb{R}^{64}$. The decoder receives both $\mathbf{z}$ and $\mathbf{d}_h$. Separate decoders initialize the shared and individual hidden states, and a cross-day module smooths each member's context across the week. The two streams then exchange information through

$$\mathbf{h}^{s,1} = \text{CrossFuse}\big(\mathbf{h}^{s,0}, \mathbf{h}^{i,0}\big), \quad \mathbf{h}^{i,1} = \text{CrossFuse}\big(\mathbf{h}^{i,0}, \mathbf{h}^{s,0}\big), \tag{8}$$

where $\text{CrossFuse}(\mathbf{q}, \mathbf{k}) = \text{LN}\big(\mathbf{q} + \boldsymbol{\sigma}(\mathbf{G}[\mathbf{q}; \mathbf{k}]) \odot \mathbf{V}\mathbf{k}\big)$ applies a learned gate to the information transferred between streams. Stream-specific heads produce the two output tensors. An activity-conditioned duration head uses learned activity embeddings and corpus-based duration priors, allowing duration errors to update the activity logits. Supplementary Section SIV gives the full encoder and decoder

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

parameterizations, posterior construction, cross-fusion and activity-conditioned duration-head equations, and the complete term-by-term objective.

Training objective. The CVAE minimizes

$$\mathcal{L}_{\text{CVAE}} = \mathcal{L}_{\text{rec}} + \beta(t)\,\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{struct}}, \tag{9}$$

where $\beta(t)$ is linearly annealed during training; the β-weighted KL formulation follows β-VAE [32]. The reconstruction loss covers activity, slot occupancy, absolute and gap-encoded timing, activity-weighted duration, and shared-event participation. The structural loss includes

$$\mathcal{L}_{\text{coord}} = \frac{\sum_{j,m,d,k} w_{j,m,d,k}\,\text{ReLU}\big(o^{\cap}_{j,m,d,k}\big)}{\sum_{j,m,d,k} w_{j,m,d,k} + \varepsilon}, \tag{10}$$

which penalizes expected overlap between a shared event and the individual activities of its participants. The weights combine day, participation, and slot-activation probabilities. Additional terms penalize negative gaps, constrain weekly Work and School days and hours by employment status, align Work starts with the declared shift, and calibrate survey targets on the validation split.

Generation. At generation time, the model samples $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and decodes both streams. Individual activity slots are retained according to the activity-gate threshold, while shared-slot participants are selected using the participation threshold. A fixed deployment post-processing pipeline then resolves residual shared–individual conflicts and applies duration caps, start-time filtering and jitter, Work and School day rebalancing, and a coordination top-up that resamples participant sets and adds shared events toward the corpus coordination targets. The gap representation prevents overlap among a member's decoded individual-stream slots; residual conflicts between the individual and shared streams are handled by the deployment pipeline. Supplementary Section SV documents the thresholds, calibration operations, and their application counts, and provides a separate raw-versus-post-processed diagnostic of their effect on household coordination. The resulting schedule S_h, together with the anchor zones from Section 4.2, conditions Level III in Eq. 1.

4.5. Level III: Trajectory completion with HEARTH

Level III implements the trajectory factor in Eq. 1. Its input is a partially dictated person-week in which Level II fixes Work, School, and shared household events at exact times. A two-token prefix supplies the home zone and one primary mandatory zone. These inputs dictate roughly one third of the week. HEARTH generates the remaining activities and locations, including transitions between home and excursions. It also conditions on zone geometry, place affordances, a demographically sampled persona, and mobility patterns learned from GPS data.

HEARTH reads the dictated week on a dense 674-token grid and generates the completion as a sequence of excursions. The grid gives the encoder an absolute time for every position. The decoder generates excursions away from home, while home periods are recovered from the dwell between consecutive excursions. A vector-quantized persona code selects one of 64 learned behavioral styles and conditions the full week. At deployment, demographics determine the distribution from which this code is sampled. Figure 3 shows the data flow.

4.5.1. Bin-grid encoder

HEARTH's encoder is a bidirectional Transformer over the 674-position weekly sequence of Section 3. Each position can attend to context before and after it, which is useful when a missing interval lies between two dictated events. Time-of-day and day-of-week indices provide the daily and weekly position of every bin. Missing activity and location fields use `[MASK]` and `[MASK_LOC]`, respectively; supplied fields retain their dictated tokens.

Each input position is embedded as

$$\begin{aligned}\mathbf{h}_t^{(0)} = {} & \mathbf{E}_a(a_t) + \mathbf{E}_z(z_t) + \mathbf{E}_\tau(\tau_t) + \mathbf{E}_\delta(\delta_t) + \mathbf{E}_r(r_t) \\ & + \alpha_{\text{mds}}\mathbf{W}_{\text{mds}}\mathbf{c}_{z_t} + \alpha_{\text{poi}}\mathbf{W}_{\text{poi}}\mathbf{q}_{z_t},\end{aligned} \tag{11}$$

where the five tables $\mathbf{E}_a, \mathbf{E}_z, \mathbf{E}_\tau, \mathbf{E}_\delta, \mathbf{E}_r$ are learned embeddings over the activity, zone, time-of-day, day-of-week, and role vocabularies, sharing a common hidden dimension. Two zone-side terms inject the geographic and semantic evidence channels. The MDS term $\alpha_{\text{mds}}\mathbf{W}_{\text{mds}}\mathbf{c}_{z_t}$ projects a low-dimensional multidimensional-scaling coordinate [33] computed from the inter-zone distance matrix, giving the model a notion of geographic proximity that varies smoothly between neighboring zones. The POI affordance term $\alpha_{\text{poi}}\mathbf{W}_{\text{poi}}\mathbf{q}_{z_t}$ projects an empirical activity-affordance distribution $\mathbf{q}_{z_t} \in \Delta^{V_a-1}$ constructed once, offline, from a regional POI dataset [34] in which every POI is mapped to one of the 11 activity classes by LLM-based classification [35]: per-zone POI counts are Laplace-smoothed ($\alpha_0 = 10^{-3}$) and normalized, so sparse zones degrade gracefully to a uniform prior. The projection $\mathbf{W}_{\text{poi}}$ is zero-initialized so the POI signal enters as an explicit, interpretable inductive bias. POI affordances are particularly informative for low-frequency activities such as Healthcare, Religious, and EatOut, whose zone preferences concentrate in specific land-use types and whose training signal in the GPS data is sparse. Special zone tokens receive zero $\mathbf{c}_z$ and $\mathbf{q}_z$ so the priors reduce to the learned discrete embedding for unobserved bins.

The encoder uses Transformer blocks [36] in a pre-norm configuration, with rotary position embeddings [37]. A separate cross-attention module keeps the home zone and one primary mandatory anchor available throughout the sequence. The mandatory-anchor token stores the Work zone whenever the resident has a Work assignment and otherwise stores the School zone. Consequently, when a resident both works and studies, the anchor memory contains Home and Work. The School location is not discarded: every dictated School bin carries its Level-I School zone at the corresponding scheduled time, and constraint-preserving decoding retains those activity–location entries unchanged. Non-prefix positions query the resulting two-token memory

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

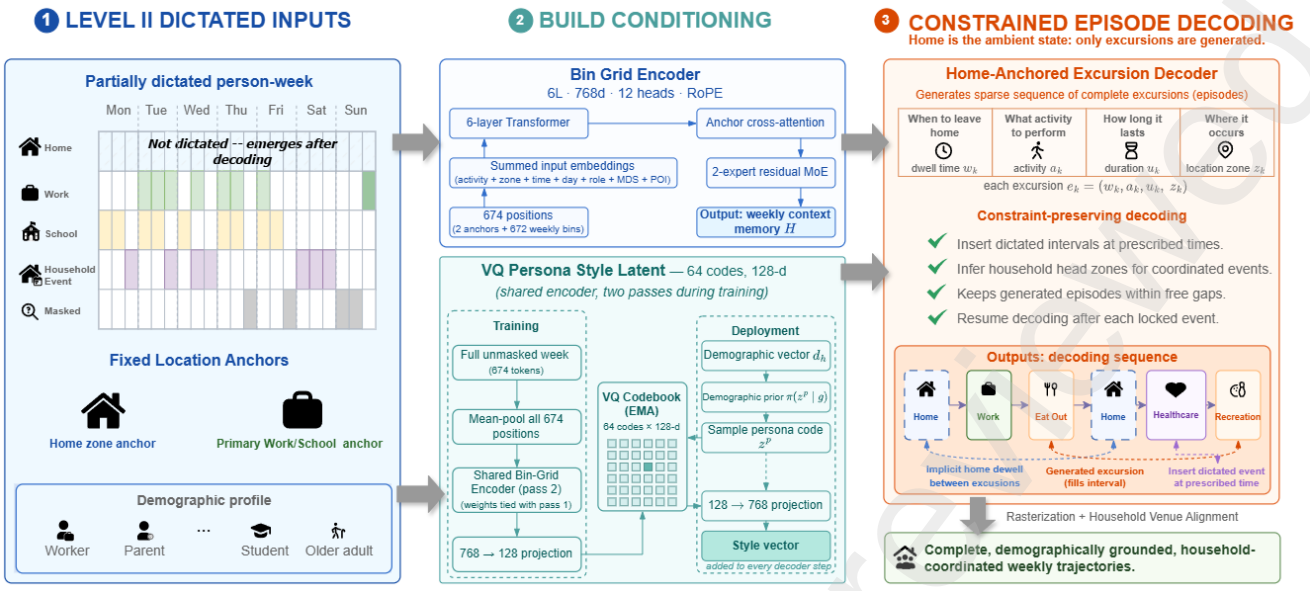

Figure 3: HEARTH (Level III): read on the grid, write in episodes. (a) The encoder and persona latent form the weekly context; the decoder predicts excursion timing, activity, duration, and zone. (b) Constraint-preserving decoding preserves all dictated fields, restricts generated excursions to intervals not occupied by dictated events, and rasterizes the completed episode sequence onto the weekly grid with Home in every unoccupied bin.

formed from anchor roles and projected anchor-zone MDS coordinates:

$$\mathbf{H}_{\text{seq}} \leftarrow \mathbf{H}_{\text{seq}} + \text{CrossAttn}(\mathbf{H}_{\text{seq}}, \mathbf{K}_{\text{anc}}, \mathbf{V}_{\text{anc}}). \tag{12}$$

A two-expert residual mixture-of-experts layer [38] then routes each position between an anchor-oriented expert and a discretionary-activity expert. This architectural separation lets the model specialize routine anchor-linked and more variable discretionary contexts. The final memory $\mathbf{H} \in \mathbb{R}^{674\times768}$ supplies cross-attention context to the excursion decoder and, after pooling, the persona latent. Section 7.4.1 evaluates these components.

4.5.2. Encoder pretraining

The encoder is first pretrained on the GPS corpus as a masked-reconstruction model [39]. A role-aware collator creates completion tasks from observed weeks. Activity masking removes all occurrences of selected discretionary activities, home-episode masking removes complete home periods, and period masking removes long time windows. Short random spans fill the remaining mask budget, whose target ratio rises from 0.3 to 0.7 during training. An activity head and an activity-conditional zone head are trained with their classification losses plus activity-marginal, activity and zone compatibility, transition, and router-supervision regularizers. The resulting checkpoint initializes HEARTH. Supplementary Section SVI gives the complete pretraining objective, loss weights, masking probabilities, and operator settings, while Section 7.4.1 measures the effect of pretraining.

4.5.3. Persona latent

Weekly behavior is person-shaped: a night-shift worker, a retiree, and a student can share a neighborhood and still live entirely different weeks. The persona latent equips HEARTH with a bank of 64 discrete behavioral archetypes, learned by vector quantization [40], and assigns every generated person one code that conditions their entire week. The encoder itself is the latent source, run twice with shared weights during training. Pass one reads the masked week and produces the decoder memory. Pass two reads the full unmasked week, mean-pools the 674 positions, and projects the pooled state through a $768 \rightarrow 128$ bottleneck into an EMA-updated codebook of 64 codes (decay 0.99, commitment weight 0.25, straight-through estimator); the selected code is projected back to 768 dimensions and added to every decoder step as a style vector. Training therefore infers each real person's archetype from that person's own week.

Deployment reverses the direction of inference because the complete week is then unavailable. Let $\boldsymbol{\phi}_k \in \Delta^{10}$ be the 11-component activity-share profile of GPS training weeks assigned to code k, and let $\mathbf{t}_g$ be the corresponding survey target for demographic group g. We fit the deployment prior by

$$\pi(\cdot \mid g) = \arg\min_{\mathbf{w}\in\Delta^{K-1}} \left\| \sum_k w_k \boldsymbol{\phi}_k - \mathbf{t}_g \right\|_2^2. \tag{13}$$

Each synthetic person n then draws $z_n^{\text{p}} \sim \pi(\cdot \mid g(n))$, where $g(n)$ is the person's survey demographic group derived from the synthetic household record. Thus, the codes are discovered from anonymous GPS behavior and connected to

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

demographics using group-level survey moments; no linked individual trajectory–demographic records enter the fit. This prior should be interpreted as group-level demographic conditioning rather than individual demographic prediction.

4.5.4. Excursion decoder

The decoder generates only excursions away from home. Home occupies the dwell between consecutive excursions. This representation follows the organization of a week into a small number of outings and reduces a 672-bin week to at most 40 decoder steps. It also makes excursion boundaries and durations explicit model outputs. The decoder is a 4-layer, 768-dimensional, 12-head Transformer. Causal self-attention processes earlier excursions, and cross-attention reads the encoder memory. At step k, the input combines the preceding excursion's activity, duration, zone, and home dwell with the current time, day, excursion index, and persona code.

Four heads predict the next excursion. The **dwell** head predicts whether the excursion is chained and models home dwell with a 24-component log-normal mixture over 1–672 bins. The **duration** head uses another 24-component mixture over 1–96 bins. The **activity** head produces logits over the 15-token input vocabulary, but Home and all four structural tokens are excluded from excursion materialization, leaving the ten out-of-home activity classes. The **zone** head produces logits over the 514-token zone vocabulary, the 512 deployment zones plus the two reserved location tokens; both reserved tokens are masked during excursion decoding, so every materialized excursion carries a real zone, and no special location token appears in a completed trajectory. These mixtures provide a flexible parameterization for multimodal timing distributions. Chained tours use zero intermediate home dwell, allowing several activities before the person returns home.

4.5.5. Training

HEARTH is trained end to end on 101,053 GPS person-weeks. The pretrained encoder is fine-tuned at 5×10^{-5}, while the decoder, persona module, and prediction heads are trained from scratch at 3×10^{-4}. We use AdamW with weight decay 0.05, a cosine schedule, bf16 autocast, 20 epochs, batch size 16, and two-step gradient accumulation. An excursion collator converts each observed week into the home-ambient sequence and supervises up to 24 excursions. The training corpus contains a mean of 13.6 excursions per person-week (median 13; 99th percentile 24); only 0.85% of weeks exceed the 24-excursion supervision limit, and 0.015% exceed the decoder's 40-excursion limit. Because the observation window can end during a home stay, the final step uses a censored dwell-survival target. The objective combines activity cross-entropy, duration and dwell mixture likelihoods, censored survival, zone cross-entropy, and VQ commitment. The cross-entropies are unweighted to avoid shifting population activity marginals. Training takes about 880 seconds per epoch on one 46 GB GPU. Across three seeds, average deployment JSD differs by at most 0.003.

Table 2
Core HEARTH configuration for the headline realism comparison. The partial input retains all supplied fields, decoding restricts generated excursions to intervals not occupied by dictated events, and rasterization assigns all remaining bins to Home. Survey steering and household venue alignment are disabled.

Item	Configuration
Weekly sequence	674 positions: 2 anchors + 672 time bins
Time resolution	15 minutes
Activity vocabulary	11 classes + 4 structural tokens
Grid-zone vocabulary	512 zones + 2 input special tokens
Encoder	768 dimensions, 6 layers, 12 heads, RoPE
Spatial encoding	16-dimensional MDS + POI affordance ($\alpha_{\text{poi}} = 0.5$)
Persona latent	VQ-EMA, 64 codes × 128 dimensions
Excursion decoder	768 dimensions, 4 layers, 12 heads, ≤ 40 steps
Timing heads	24-component log-normal mixtures
Decode-time consistency	clamping + rasterization

Table 2 consolidates the core HEARTH architecture and the decode configuration used in the headline realism comparison.

4.6. Component Taxonomy

For evaluation, we separate learned generation, constraint-preserving operations, and framework post-processing. The learned components are the population synthesizer, LLM-grounded corpus construction, CVAE coupling, HEARTH encoder and decoder, and VQ persona. At Level II, teacher checks and the fixed deployment cleanup impose schedule consistency. At Level III, constraint-preserving decoding keeps dictated fields fixed, restricts generated excursions to intervals not occupied by dictated events, and rasterizes the completed episode sequence with Home in every remaining bin. Demographic steering adjusts activity logits through the survey log-ratio bias described below. Household venue alignment copies the head-inferred coordinated-event zone to the other participants after trajectory completion.

Constraint-preserving decoding. Work and School enter decoding as fixed activity–time episodes at their Level-I anchor zones. Coordinated events enter with fixed activity, time, and participants but no location. Constraint-preserving decoding leaves all of these supplied fields unchanged. HEARTH infers each coordinated event's zone from the household head's trajectory context and generates additional excursions only within intervals not occupied by dictated episodes. For each generated excursion, the dwell head selects its start, the duration head selects a length that ends before the next dictated episode, and the activity and zone heads select its purpose and destination. Rasterization then combines the dictated and generated episodes on the 672-bin grid and assigns every unoccupied bin to Home.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

Table 3
Data source characterization: unit, resolution, labels, the level each source trains or calibrates, and the reliability assumption that use builds on.

Source	Unit	Resolution	Labels	Role	Reliability assumption
Census	aggregate cell	regional / demographic	household and person attributes	population constraints (Level I)	authoritative marginals
Time-use survey	respondent / group	national / diary time bin	activity and joint time	demographic behavior targets (II, III)	authoritative group statistics; no spatial claim
GPS corpus	anonymous person-week	metro / 15 min / zone	inferred activity + location	mobility dynamics (II, III)	strong spatiotemporal evidence; no demographic claim
POI / geography	place / zone	metro / point	semantic category	affordance and geometry priors (III)	semantic evidence only
LLM teacher	household prompt	synthetic / event	relational proposal	semantic coordination (II)	filtered, calibrated proposal; never ground truth

Household venue alignment. After trajectory completion, framework post-processing copies each head-inferred coordinated-event zone to the event's other participants while leaving activity, time, and membership unchanged. Household venue alignment is not learned and requires no additional training.

Demographic steerability. Demographic steering supports applications that require closer agreement with demographic targets. The **survey-conditional log-ratio bias** adjusts activity logits using the survey distribution $\pi_g^{\mathrm{TUS}}(a \mid \delta_\star, \tau)$ and the model's unconditional slot marginal $p^{\mathrm{ref}}(a \mid \delta_\star, \tau)$ from a cached unbiased decode:

$$b_{g,a}^{\mathrm{TUS}}(\delta_\star, \tau) = \kappa_g \log \frac{\pi_g^{\mathrm{TUS}}(a \mid \delta_\star, \tau) + \epsilon}{p^{\mathrm{ref}}(a \mid \delta_\star, \tau) + \epsilon}, \tag{14}$$

The bias is clipped to $[-5, 5]$ and added to each person's logits before the softmax. If the survey and model distributions have the same support and the logits match the unconditional marginal, $\kappa_g = 1$ maps that marginal to the survey conditional. Because decisions across slots are coupled in practice, we treat κ_g as a steering strength and select it per group by grid search over $\kappa_g \in \{1, 2, 3, 4, 6, 8, 12\}$, decoding a stratified sample of 500 synthetic residents per group and choosing the value that minimizes that group's activity-share divergence from its survey target. The bias retains time-of-day structure and can therefore adjust group-specific timing patterns. The survey covers people aged 15 or older, so younger residents are excluded from steering and from the reported steerability results, although the generator still produces schedules for all ages. Because the survey has no Student employment category, student groups use the not-working target for the same age band. The 16 covered strata account for roughly half of the deployed person-weeks.

5. Experimental Design

5.1. Sources and preprocessing

Table 3 characterizes the five evidence sources, and the paragraphs below give their concrete instantiations for the Tokyo metropolitan deployment.

GPS mobility corpus. HEARTH is trained and evaluated on anonymized person-week GPS data from the Tokyo metropolitan area. After conversion to the common 672-bin weekly grid, the corpus contains 101,053 person-weeks for training, 12,632 for validation, and 12,632 for testing. The split is person-disjoint, so all weeks from the same individual remain in one partition. Zones are observed in 90.16% of the training bins, and the corpus contains no demographic labels. We derive activity–zone sequences using two components of our prior work. Raw trajectories are first segmented into staypoints and labeled through weakly supervised dual-path activity inference [41]. POI semantics are then mapped to the activity vocabulary using LLM-based classification [35]. Only labels above the calibrated confidence threshold supervise the model; unresolved bins receive `[UNK]`. The retained staypoints are rasterized to 15-minute bins and assigned to the 512-zone partition.

Census marginals. Synthetic households are sampled to match regional census marginals and conditional tables from e-Stat, the Portal Site of Official Statistics of Japan [42], over the eight demographic axes of Section 4.2.

Time-use survey. Per-demographic activity targets and dyad coordination budgets derive from the 2016 Japanese national time-use survey (JTUS-2016) [43], aggregated onto the model's 11 activity classes.

Points of interest. Zone-level activity affordances are computed from a regional POI dataset [34], each POI mapped to an activity class by LLM-based classification (Section 4.5.1).

Mobility reference. The mandatory-location assignment of Level I is evaluated against Veraset-derived mobility data [41]. Unless otherwise noted, all distributional comparisons use Jensen–Shannon divergence (JSD); lower is better.

5.2. Evaluation protocol and metrics

The trajectory-level evaluation follows the deployment workflow. HEARTH trains on the 101,053-week GPS training split and completes 395K dictated person-weeks from one shard of the 1.98-million-week synthetic population. Survey steering and household venue alignment are disabled for this realism comparison. We score 50,000 sampled completions against the held-out GPS test split. On one NVIDIA L40S GPU, decoding 50K weeks takes approximately 18 minutes. Realism is measured by six Jensen–Shannon divergences (JSDs) between generated and held-out reference

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

distributions using shared fixed-width histogram bins. DailyLoc is the mean of seven day-of-week-specific JSDs for distinct zones visited per person-day; Departure compares episode start times; Duration is the mean of the eleven activity-specific episode-duration JSDs; Distance compares individual trip distances rather than daily totals; Radius compares weekly radii of gyration; and Category compares activity-class shares. Their unweighted mean is the Std-6 score. The OD metric is the JSD between generated and reference 512×512 origin–destination transition matrices, each flattened and normalized by its global total, so the comparison is a single divergence over all origin–destination pairs and weights each pair by its share of total flow.

All nine comparison systems are fitted or trained on the same GPS training split, receive the same dictated deployment inputs, and are converted to the common 672-bin, 11-activity, and 512-zone representation for evaluation. Section 7.3 reports the model-specific adaptations required by this common task.

5.3. System variants under test

The evidence of Section 6 comes from four controlled variants, each changing one source or alignment mechanism while holding the rest of the deployed system fixed:

- *Unconditional persona* replaces $\pi(z^{\mathrm{p}} \mid \mathbf{d}_h)$ with a uniform distribution over the 64 codes, removing demographic survey evidence from persona selection.
- *POI withheld* zeros the POI-affordance input at inference while retaining zone geometry and the learned zone representations.
- *No household-schedule latent* removes all Level-II dictated events but retains the Level-I home and work/school anchors. The trajectory model therefore receives demographics and anchors but no household schedule.
- *GPS-dictated oracle* replaces generated upstream schedules with dictated views constructed from held-out GPS weeks. The collator retains recurring Work and School episodes and the anchor prefix, masks the remaining week, and scores the completions against the held-out split.

The first three variants decode the full 395K-week deployment shard. The oracle uses all 12,632 held-out GPS weeks. Every result is produced by the same evaluation pipeline.

6. System-Level Evaluation

Five research questions structure the evaluation. **RQ1:** can unlinked aggregate and anonymous sources be combined into a synthetic population that stays consistent across population, household, demographic, temporal, and spatial levels (Section 6.1)? **RQ2:** what does each evidence source contribute, measured by removing or weakening it while the rest of the system holds (Section 6.2)? **RQ3:** does the hierarchical latent organization outperform alternative couplings, tested at its most load-bearing link, the household-schedule latent, through removal at the trajectory level (Section 6.4) and through per-member, late-fusion, and flat joint-decoding alternatives at the schedule level (Section 7.2)? **RQ4:** how does the system behave when a source is withheld at inference, the evaluation evidence shifts, or generated inputs are replaced by observed ones (Sections 6.3–6.5)? **RQ5:** do the individual levels remain competitive with specialized baselines on their native tasks (Section 7)?

6.1. End-to-end cross-level consistency

The first question the system must answer is whether the properties each source supports survive into the final object together. Three cross-level axes are measured on the deployment decode, each connecting a source's authoritative level to generated behavior two levels away, and Table 4 assembles the conditional-fidelity matrix that extends them.

Demographic evidence reaches behavior. Across the 16 covered groups, the mean JSD between generated activity-time proportions and survey targets is 0.0406 (Table 5). Worker groups show the closest agreement, with JSDs from 0.0102 to 0.0315. Young student groups fall in the middle, while nonworker groups have the largest residuals, reaching 0.0845. The two mid-aged student groups are exceptions to this ordering: their JSDs are 0.0500 and 0.0792, and they also have the smallest samples (n=425 and 390). More generally, groups with dictated Work or School schedules provide a stronger scaffold for demographic steering than groups whose weeks are mostly discretionary. Because the steering targets are applied directly, Table 5 measures constraint satisfaction. The unsteered demographic result in Table 6 measures generalization.

Household structure reaches shared placement. Without venue alignment, independently completed trajectories co-locate at shared events with probability 0.026. Alignment raises co-presence to 0.994, while Std-6 changes from 0.0252 unaligned to 0.0256 aligned under the corpus-wide reference. Level II supplies the coordinated skeleton and HEARTH the head's event zone; framework post-processing supplies near-exact co-presence at negligible realism cost.

Geography reaches flows. HEARTH achieves an OD-flow JSD of 0.3007, the lowest among the nine baselines. Together with the home-conditioned destination result in Table 4, this shows that the upstream spatial context remains visible not only in individual trip distances but also in aggregate origin–destination structure.

Table 4 evaluates five conditional relationships. Employment-conditioned activity divergence rises from 0.026 for workers to 0.128 for nonworkers, matching the survey-fidelity gradient. Destination choice remains home-conditioned, with population-weighted per-origin JSD 0.0916 across 64 clusters. The Level-II schedule preserves a six-hour separation between Day- and Evening-shift Work starts. Child-conditioned care is 3.2× more common in the dictated layer, but falls to 1.09× end to end as free generation adds patterns learned from GPS without household labels. Coordination still increases monotonically with household

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

Table 4
End-to-end cross-level conditional fidelity on the shipped deployment decode: each row conditions generated behavior on an upstream attribute introduced one or two levels earlier. References: survey conditionals for employment; the dictated schedule for shift, children, and household size (the GPS corpus carries no such labels); the training-split GPS empirical conditional for home zone, a conditioning reference that requires per-origin mass.

Conditional relation	Metric	Value	Reading
Employment → activity time-share	cond. JSD vs survey (Worker/Student/Nonworker)	0.026 / 0.067 / 0.128	worker–nonworker gradient of Table 5 reproduced
Work shift → Work start	median generated start (Day/Evening)	08:45 / 14:45	shift timing transfers; Overnight spans midnight
Child presence → ChildCare+DropOff	with/without-children hours ratio	3.2× dictated; 1.09× end-to-end	separation carried by the schedule latent
Household size → shared hours	per-capita scaling by size {2,3,4+}	10.1 / 12.8 / 14.3 h, monotone	size gradient preserved; decode/dictated ratio stable (≈2.5×)
Home zone → destinations	weighted mean per-origin JSD (64 origin clusters)	0.0916	conditioning holds at all 64 origins; coarser origin binning contributes to the level

Table 5
Stratified steerability of the steered deployment: JSD between generated activity-time proportions and TUS group targets, all 16 covered strata.

Group	n	JSD
Mid-Aged Male Worker	56,370	0.0102
Elderly Male Worker	10,324	0.0125
Mid-Aged Female Worker	49,475	0.0178
Young Female Worker	3,036	0.0202
Young Female Student	5,003	0.0205
Young Male Student	7,758	0.0207
Young Male Worker	3,778	0.0222
Young Male Nonworker	827	0.0287
Young Female Nonworker	664	0.0299
Elderly Female Worker	10,722	0.0315
Mid-Aged Female Student	425	0.0500
Elderly Male Nonworker	17,692	0.0672
Mid-Aged Female Nonworker	5,201	0.0721
Mid-Aged Male Student	390	0.0792
Mid-Aged Male Nonworker	6,015	0.0819
Elderly Female Nonworker	22,293	0.0845
Stratum mean		**0.0406**

size. Thus, the schedule carries household and demographic structure, while the free layer needs better participation calibration. It raises nonworker Work participation from 0.04 to 0.15, consistent with Table 5.

6.2. Source contribution

Removing a source should damage the properties that source supports and little else; a system whose sources were interchangeable would have needed no source-specialized architecture. Table 6 tests this by removing one evidence source at a time from the deployed system's inference path, decoding the full shard, and scoring three columns that live at different levels: aggregate weekly realism (Std-6), demographic-conditional time use (the 16-stratum mean of Table 5's metric, measured without survey steering), and OD flow.

The persona prior primarily supplies demographic separation. Replacing the demographic prior with a uniform

Table 6
Source-contribution ablation: one evidence source removed from the inference path per row, full 395K-week decode, identical evaluation chain. Std-6 and OD flow are scored against the held-out GPS test split; the demographic column is the 16-stratum mean JSD of generated activity-time proportions against survey targets, measured without survey steering.

Variant	Std-6	Demographic	OD flow
Full system (realism config)	0.0252	0.0481	0.3007
Unconditional persona (survey prior removed)	0.0271	0.0982	0.3128
POI withheld at inference	0.0252	0.0478	0.3007

distribution changes Std-6 only from 0.0252 to 0.0271 and OD divergence from 0.3007 to 0.3128. In contrast, demographic divergence nearly doubles, from 0.0481 to 0.0982. This separation of effects matches the intended source routing. GPS dynamics and the dictated schedule govern aggregate realism, while the survey-conditioned persona prior differentiates behavior across demographic groups. Because this prior is the only survey input active in the unsteered realism configuration, the uniform-persona experiment also represents removal of survey evidence at inference. Applications that additionally enable the survey log-ratio correction achieve the 0.0406 stratified mean reported in Table 5.

The POI-withholding experiment isolates inference-time dependence rather than the independent training-time contribution of POI evidence. Zeroing the affordance input at decode time leaves Std-6, demographic divergence, and OD flow unchanged (0.0252 / 0.0478 / 0.3007), with no measurable shift in the POI-adjacent granular metrics. The shipped decoder therefore does not require access to the POI channel at inference. Because the encoder ablation in Section 7.4.1 removes POI together with the other spatial components, it establishes the combined value of the spatial encoder bundle but does not identify the training-time effect of POI alone.

6.3. Robustness to withheld and shifted evidence

Two experiments test robustness to changes in available evidence. First, rescoring deployment outputs against the complete 126,317-week GPS corpus, including the training split, changes every Std-6 average by less than 0.0001

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

Table 7
Architecture ablation: the household-schedule latent removed from the inference path, full 395K-week decode scored against the held-out GPS test split. The no-latent variant contains no shared events, so shared-event co-presence has no support. The cross-level conditional relations of Table 4 extend this comparison.

Variant	Std-6	Demographic	OD flow	Co-presence
Full hierarchy	0.0252	0.0481	0.3007	0.994
No household-schedule latent	0.0237	0.0760	0.2743	undefined

relative to the held-out reference. The reported ranking is therefore not sensitive to this reference shift. Second, the POI-withholding result of Table 6 also serves as the missing-source test. Other missing-source cases have explicit fallbacks. The trained CVAE requires no LLM access, the persona prior can use pooled national statistics when a local survey is unavailable, and absent census joint tables revert to the sequential conditional chain. These fallbacks concern inference availability; the corresponding sources remain part of training where specified.

6.4. Architecture ablation: the household-schedule latent

The hierarchy's most load-bearing link is the household-schedule latent, the alignment variable through which all Level-II evidence reaches trajectories. The cleanest test removes it: the no-latent variant hands the trajectory operator only demographics and spatial anchors, and the model generates each member's entire week free-form. Table 7 reports the outcome.

Removing the household schedule preserves aggregate marginals but removes cross-level dependence. Without dictated Level-II events, Std-6 improves slightly from 0.0252 to 0.0237 and OD divergence also falls. The unconstrained GPS-trained decoder can match population-level distributions more freely when it no longer has to preserve fixed events across one third of the week. Conditional behavior moves in the opposite direction. Demographic divergence increases by 58%, from 0.0481 to 0.0760, because Work, School, and shift patterns are no longer supplied by the schedule. The model instead reproduces these behaviors at their population-average rates. Because the ablation removes all shared events, co-presence is undefined; the schedule latent therefore supplies demographic and household dependence rather than lower unconditional marginals.

6.5. Error propagation across levels

The last question is where residual error originates: in the trajectory operator itself, or in the generated inputs it consumes. Table 8 brackets the answer with two operating points. The deployment point completes generated schedules for generated households. The oracle point replaces both with reality: dictated views built from held-out real GPS weeks through the training-time anchors-only view, completed by the same model under the same configuration and scored against the held-out split itself.

Table 8
Error propagation: identical model and decode configuration, inputs varied, both rows scored against the held-out GPS test split.

Operating point	Inputs	Reference	Std-6	OD
Deployment	generated pop. + schedules	held-out 12.6K	0.0252	0.3007
GPS-dictated oracle	real dictated views	held-out 12.6K	0.0146	0.2387

With observed dictated inputs, HEARTH reaches a Std-6 score of 0.0146, compared with 0.0252 under deployed synthetic inputs. The 0.0106 gap reflects the combined effects of population sampling, schedule generation, and the shift from GPS-shaped to CVAE-shaped dictation; this experiment does not identify their individual contributions. Because dictated bins are copied into the output and constitute roughly one third of each week, the oracle is a favorable diagnostic operating point rather than a deployable performance estimate. Holding the trajectory operator fixed, the improvement nevertheless localizes a substantial share of the remaining end-to-end error to the upstream inputs and motivates further refinement of population and schedule generation.

7. Module-Level Evaluation

System-level evidence established, we ask whether each level holds its own against specialists. Each level is validated against the references its sources define: the population against census targets, the teacher corpus and schedule model against survey and GPS references, and the trajectory operator against nine baselines on the deployment task.

7.1. Level I: Demographic and spatial population synthesis

Level I integrates regional aggregate demographic evidence with mobility-derived spatial evidence to construct the population-level alignment layer of PopTraj. It produces two linked outputs: household demographic records that condition the Level-II schedule generator and demographic behavior, and home–mandatory-activity anchors, which provide the fixed spatial context for Level-III dictated weeks. We evaluate these outputs against their corresponding reference data: aggregate census tables for demographic structure and mobility observations for spatial anchors.

The proposed population generation method is compared with four sample-free population-synthesis methods that generate populations directly from aggregate distributions: **independent marginal sampling (IMS)**, which draws each member attribute independently from its marginal distribution; **iterative proportional fitting (IPF)** [44]; the **sample-free reconstruction method (B&T)** [45], which estimates household-type joint distributions from aggregate tables and assembles households from an individual pool; and **GenSynthPop (GSP)** [20], which samples attributes sequentially from conditionals on previously assigned attributes. All methods generate the same eight demographic attributes, with unmodeled attributes sampled from their

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

Table 9
Population synthesis evaluation on $N = 100{,}000$ synthetic Tokyo households against e-Stat targets.

	Proposed	IMS	IPF	B&T	GSP
Marginal fidelity (standardized RMSE, SRMSE)					
Household size	0.004	0.005	0.006	**0.003**	0.006
Age | male	<**0.001**	0.013	0.343	0.042	0.013
Age | female	<**0.001**	0.016	0.308	0.123	0.015
Employment status	0.004	**0.003**	0.235	0.017	**0.003**
Income percentile	**0.011**	0.086	0.017	0.014	0.075
Religious affiliation	0.003	0.003	**0.002**	**0.002**	0.003
Pairwise joint-distribution fidelity (JSD)					
Size × head sex	<**0.001**	0.054	0.001	0.001	0.031
Size × sex × head age	0.012	0.165	**0.001**	0.122	0.082
Age × emp. | male	<**0.001**	0.227	0.070	0.013	0.001
Age × emp. | female	<**0.001**	0.213	0.062	0.007	0.001
Size × income	<**0.001**	0.007	0.223	0.222	0.004
Head sex × income	<**0.001**	0.030	0.198	0.196	0.011
Household structural consistency (MAE)					
Employment rate (age × sex)	**0.001**	0.167	0.023	0.039	0.014
In-school fraction (15–24)	**0.002**	0.467	0.063	0.012	0.004

Lower is better (SRMSE, JSD, MAE); bold marks the best per row.
Work shift and vehicle ownership are excluded from marginal fidelity evaluation as no Tokyo census reference is available.

corresponding aggregate marginals. Each method generates $N = 100{,}000$ households for the same Tokyo case using identical aggregate demographic inputs. The evaluation is organized according to the role of each property in downstream conditioning. Marginal SRMSE measures demographic prevalence and therefore the composition of the conditioning groups presented to later modules. Pairwise JSD measures whether specified dependencies among household and person attributes are preserved. Structural MAE measures whether age-specific employment and school participation remain consistent with the constraints used to generate work and school schedules.

The proposed method achieves the lowest marginal error on age structure and income, and lands within 0.002 of the best baseline on the remaining axes. Table 9 (upper block) reports SRMSE against e-Stat census targets. The proposed method attains SRMSE < 0.001 on both age pyramids, reflecting the post-hoc age-correction step, and leads on income percentile (0.011 vs. 0.014–0.086). IPF and B&T are competitive on religious affiliation (0.002), where those axes are explicit IPF fitting targets, and trail by one to two orders of magnitude on age (0.04–0.34), exposing the cost of high-dimensional table sparsity outside their directly fitted margins.

Sequential conditioning primarily improves household-level joint distributions. In the middle block of Table 9, the proposed method has the lowest JSD on five of six attribute pairs with census contingency tables. Relative to independent marginal sampling (IMS), it reduces JSD by 93% for the head-age structural triple and by more than 99% for both age–employment pairs. It also improves the head-age triple from 0.082 with GenSynthPop (GSP) to 0.012, a 6.8× reduction. The nine-role household taxonomy, head-relative age-gap priors, and population-level age-pyramid

Table 10
Mandatory activity location assignment evaluation on $N = 100{,}000$ synthetic Tokyo households against Veraset mobility data.

Metric	Prop.	Marg.	Rad.
Home zone marginal JSD (↓)	**0.068**	**0.068**	**0.068**
Work zone marginal JSD (↓)	0.069	**0.0025**	0.092
School zone marginal JSD (↓)	0.086	**0.0087**	0.087
Home–work distance (Wasserstein, km, ↓)	**6.441**	13.528	9.737
Home–school distance (Wasserstein, km, ↓)	**7.171**	13.995	7.669
Home–work OD matrix (CPC, ↑)	**0.360**	0.0697	0.271
Home–school OD matrix (CPC, ↑)	0.233	0.0228	**0.235**

Arrows indicate the direction of improvement; bold marks the best result in each row.
Home zones are identical across all methods.

correction supply structure absent from GSP. IPF nearly matches the structural triple it fits directly, but records JSD values of 0.20–0.22 on the income joints, showing that fitted tables do not constrain unmodeled dependencies. In the lower block, the proposed method achieves MAE $\leq$ 0.002 for employment and school participation by age; IMS deviates by roughly two orders of magnitude after breaking the age–employment relationship.

The demographic records determine which residents require mandatory activity anchors and provide the home origins for spatial assignment. We therefore assign each employed or enrolled member a Work or School location using the proposed hybrid gravity model and compare it with destination-marginal sampling and the radiation model [46]. All methods use identical synthetic home zones, assign Work and School locations separately, and estimate their destination-choice quantities from the same mobility reference. This protocol isolates destination assignment conditional on common home origins. Because these anchors enter Level III as fixed spatial context, their evaluation considers three complementary properties. Destination-zone JSD measures aggregate spatial demand, commute-distance Wasserstein distance measures travel-length fidelity, and CPC measures origin–destination structure. The assignment is evaluated against Veraset-derived mobility data [41]; the home-zone marginal, which is identical across methods, is reported as a reference.

The proposed hybrid gravity assignment provides the most balanced spatial fidelity across destination frequencies, commute distances, and OD structure. It attains the lowest Wasserstein distance for both Work (6.441,km) and School (7.171,km), and the highest Work CPC (0.360). Destination-marginal sampling achieves the lowest destination JSD by construction, but performs poorly on commute distance and OD structure because it contains no origin-conditioned distance mechanism. The radiation model improves on this baseline and slightly exceeds the proposed method on School CPC (0.235 versus 0.233), but has less accurate School distance fidelity (7.669 versus 7.171,km) and weaker Work performance. The proposed assignment is therefore the only method that jointly maintains strong destination, distance, and OD fidelity.

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

Together, the two evaluation blocks assess complementary components of the Level-I output. Demographic synthesis determines who is represented and how household attributes co-vary, thereby conditioning schedule generation and demographic behavior. Mandatory-location assignment supplies the home, Work, and School anchors that constrain the spatial context of trajectory generation. These results establish the fidelity of the inputs passed to Levels II and III; behavioral realism generated from those inputs is evaluated separately in the subsequent sections.

7.2. Level II: Teacher corpus and schedule model

We first compare the generated teacher corpus with its survey targets. OpenAI's o4-mini generated the survey-calibrated, two-pass corpus using the Chat Completions API's default decoding settings. The comparison separates quantities the prompt constrains from quantities it leaves free: agreement on the former verifies that the calibration closed, and agreement on the latter is evidence about the generator itself. Weekly Work and School hours are injected targets, and the corpus lands close to them: generated Work averages 41.0 h against the survey's 39.0 h, and School 32.8 h against 33.5 h (Fig. 4, left). Durations receive no target at all, yet the daily distributions peak at survey-typical values for EatOut (1.9 h), Errands (1.1 h), Visit (2.0 h), and Religious (0.9 h) with realistic right skew (Fig. 4, right), so these episode lengths come from the generator rather than the prompt.

The teacher corpus preserves the survey's ordering of dyad involvement—couples, child–mother, then child–father—with each generated share within three percentage points of its reference (Fig. 5). For couples, it also reproduces the within-dyad activity composition nearly exactly: generated/survey shares are 42%/41% for FreeTime, 30%/30% for UnpaidWork, and 29%/29% for PersonalCare. The agreement therefore extends beyond total joint time to both who coordinates and, for couples, what they do together. Parent–child activity composition remains the principal exception, as shown in Fig. 6.

The same conformance pattern appears across o4-mini, DeepSeek-V4-Pro, and Qwen3.6-Plus (Table 11). Recurring Work and School constraints are comparatively easy to satisfy, clearing on the first attempt for more than 92% of households, whereas coordinated-activity constraints clear for 76–89% and account for most regeneration. Best-effort retention after six attempts remains at or below 2%. This result demonstrates the portability of the conformance pipeline across generators; it does not establish equivalent corpus fidelity, which is evaluated here only for the o4-mini corpus.

The distilled CVAE preserves the teacher corpus's three central structures: demographic gradients in mandatory activity, weekly participation rhythms, and household coordination. It also corrects several teacher deviations; for example, Young Female Work decreases from 46 h in the corpus to 39 h in the CVAE, matching the survey target, and both student groups move closer to their targets (Fig. 7). Weekly

Table 11
Share of households cleared by each generation pass on each attempt (attempts 3–6 pooled; rows sum to 100%). All three generators use the identical pipeline, so the figures measure conformance cost; corpus fidelity is evaluated separately above.

Model	Attempt 1	Attempt 2	Attempts 3–6	Best-effort
Recurring pass (Work, School)				
o4-mini	93.6%	6.4%	0.0%	0.0%
DeepSeek-V4-Pro	92.0%	7.3%	0.7%	0.0%
Qwen3.6-Plus	92.7%	7.3%	0.0%	0.0%
Coordinate pass (eight shared types)				
o4-mini	76.4%	14.5%	7.3%	1.8%
DeepSeek-V4-Pro	82.0%	12.0%	4.7%	1.3%
Qwen3.6-Plus	88.7%	7.0%	4.3%	0.0%

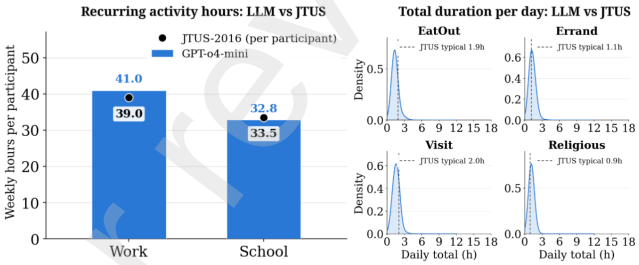

Figure 4: Recurring stream evaluation: teacher corpus vs. the time-use survey (JTUS-2016). Left: per-participant weekly Work and School hours (constraint satisfaction). Right: generated daily-duration densities for four discretionary activities with survey-typical durations marked (generalization).

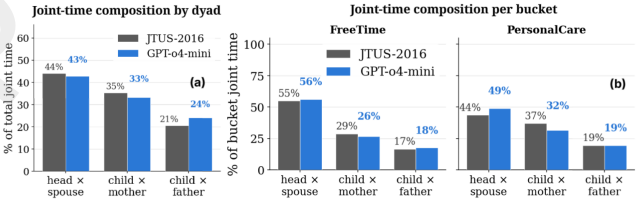

Figure 5: Coordinate stream, joint-time composition. (a) Share of total joint coordinate time per dyad; the survey rank order is reproduced within three points. (b) The same composition within the FreeTime and PersonalCare buckets.

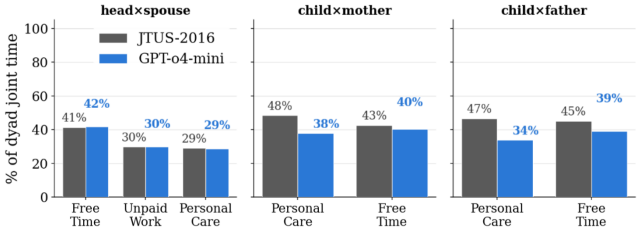

Figure 6: Coordinate stream, bucket composition within each dyad (generated vs. survey). The split is near-exact for couples; the parent–child dyads over-allocate UnpaidWork relative to the survey, a named teacher bias.

participation retains the weekday–weekend pattern, shift-conditioned Work starts retain their distinct modes, and joint activity continues to increase with household size while dyad composition remains within two to three percentage points

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

Table 12
Level-II schedule-generation benchmark on the 559-household held-out set. Best-generated pipeline per column in **bold**; dyad-order ties at 1.0 unmarked. †degenerate decode. Protocol and metric definitions in the text.

Method	Hours	Share JSD	Coord.	Dyad ρ	Ovl.	Dist.
Corpus (teacher)	3.44	0	0	1.0	0.026	0.70
CVAE (deployed)	**2.48**	**0.083**	**0.87**	1.0	**0.000**	**0.79**
GRU	3.77	0.129	18.6	1.0	0.165	0.70
LSTM	3.60	0.138	23.9	1.0	0.183	0.70
Transformer	3.74	0.129	18.7	1.0	0.168	0.70
ActVAE-style CVAE	5.25	0.168	16.5	1.0	0.113	0.69
Conditional WGAN-GP	8.03	0.249	14.4	−0.26†	0.000†	0.52†
Flat household CVAE	9.63	0.202	19.2	1.0	0.109	0.65

of the corpus (Figs. 8–10). The main remaining error is the upper tail of Mid-Aged Male Work: the CVAE produces 54 h against 60 h in the survey, reflecting the effect of the weekly-hour caps. Distillation therefore retains the principal household and temporal relationships but compresses unusually long Work schedules.

We evaluate household-schedule generation on 559 held-out households, with each pipeline generating three schedules per household for comparison against the corpus and survey references. Six measures cover survey Work/School-hour error (Hours), corpus activity-share divergence (Share JSD), coordination-volume error (Coord.), dyad rank correlation (Dyad ρ), member overlap (Ovl.), and schedule distinctness (Dist.). Supplementary Section SV documents its thresholds, operations, and operation counts, and separately isolates its effect on coordination using a 3,000-household deployment diagnostic; that diagnostic is distinct from the 559-household held-out benchmark. Table 12 compares six pipelines trained on the same split and demographic inputs: **GRU**, **LSTM**, and **Transformer** event generators; an **ActVAE-style** member-level conditional VAE [22]; a **conditional WGAN-GP** schedule generator [47], adapted to mobility generation following Feng et al. [7]; and a **flat household CVAE**. The member-level grid models infer shared events by post-hoc time matching, whereas the flat CVAE jointly decodes the household without shared/individual factorization.

Table 12 shows that per-member generators reproduce individual rhythms but do not recover household coordination: only 4–5% of declared shared events are independently confirmed, and their coordination errors range from 16.5 to 23.9, compared with 0.87 for the deployed Level-II pipeline. The flat household CVAE further shows that joint household input alone is insufficient, with coordination error 19.2 and Work/School-hour error 9.6 h, compared with 2.5 h for the proposed pipeline. These results identify the explicit shared/individual streams and participant masks as the components responsible for preserving household structure in this benchmark. After fixed deployment post-processing, the Level-II pipeline is overlap-free, achieves the highest distinct-schedule rate (0.79), and generates 43 household-weeks per second. Because no observed household-labeled weekly schedule reference is available, this benchmark measures fidelity to the calibrated teacher corpus and survey targets rather than fidelity to a real household-week distribution.

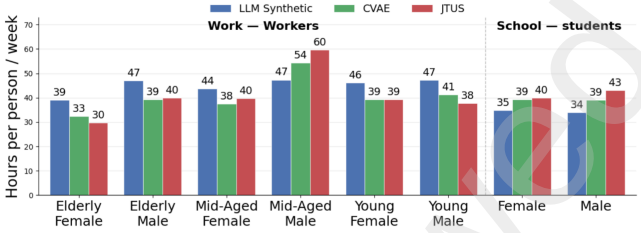

Figure 7: Per-person weekly Work and School hours by demographic group: teacher corpus, CVAE, and JTUS target.

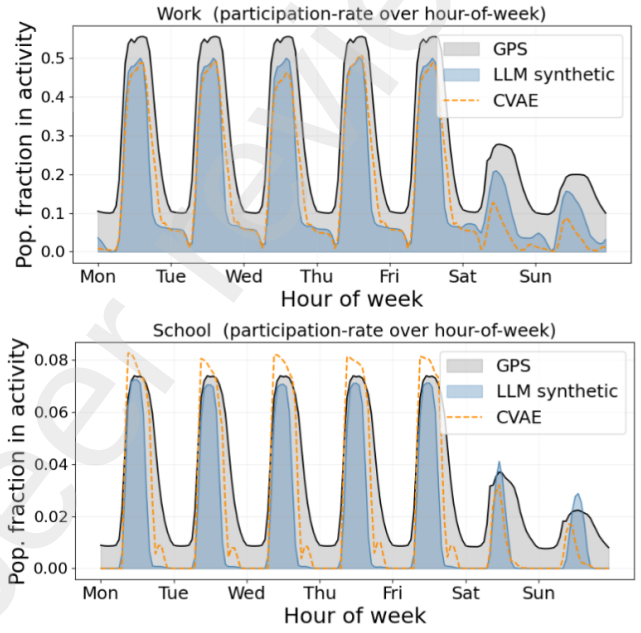

Figure 8: Weekly Work (top) and School (bottom) participation rate across the 168-hour week: GPS reference, teacher corpus, and CVAE.

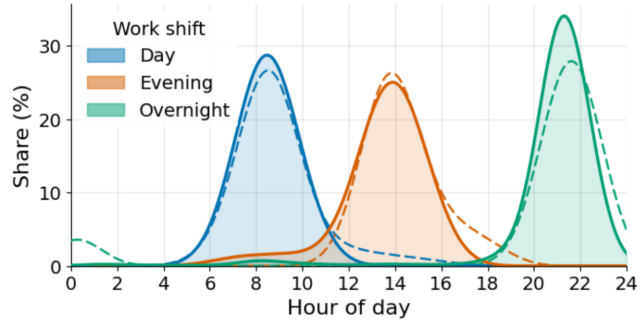

Figure 9: CVAE Work start-hour distribution by declared shift (generated solid vs. target dashed).

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

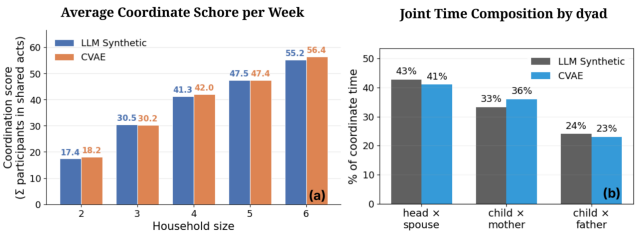

Figure 10: (a) Weekly coordination score (participant-hours in shared activities) vs. household size. (b) Joint-time composition by dyad. CVAE vs. teacher corpus.

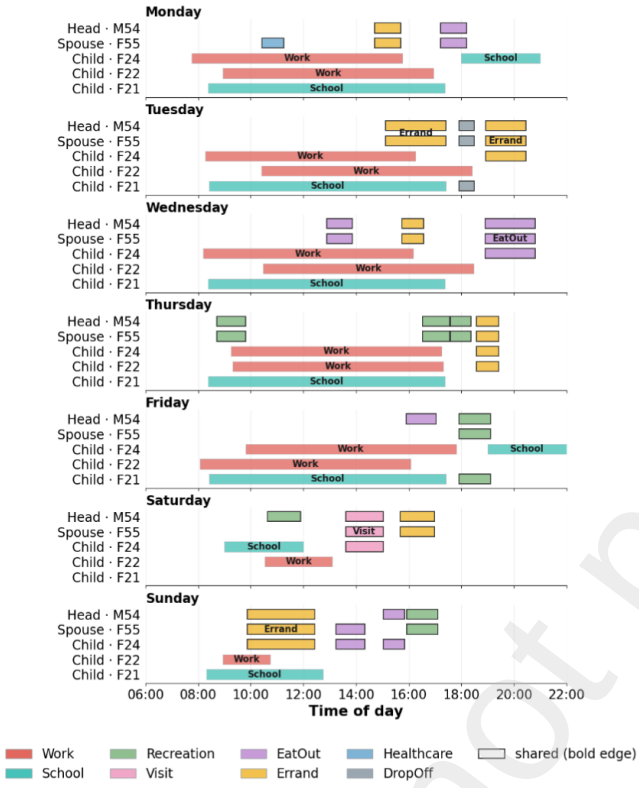

Figure 11: A generated five-member household-week: Work/School plus bold-edged shared discretionary activities.

7.3. Level III: Trajectory realism against nine baselines

HEARTH enters this comparison carrying capabilities no baseline in the panel expresses together: demographic conditioning through the persona prior, household coordination with venue-level co-presence, and survey steerability, each measured on its own axis in Section 6. The realism panel switches the levers off (steering and household pooling, for every system including ours) precisely so the comparison lands on the one ground all ten systems share, the distributional realism of completed weeks. The question Table 13 answers is whether a model that carries those capabilities pays for them in realism on the baselines' own terms.

Baseline implementations and task adaptations. Our deployment task is *dictated-week reconstruction*: each method receives fixed Work/School activity–location anchors and coordinated events with fixed activity and time but unresolved location, preserves the supplied fields, completes the missing fields and free intervals, and outputs both activity and zone. Unlike conventional sparse-trajectory recovery, no observed mobility history exists for the synthetic resident. We state every required task adaptation explicitly; the -DH suffix denotes added joint activity and zone outputs.

Full-trajectory generation, adapted to reconstruction.

- **TimeGeo** [24] is a mechanistic simulator that generates a complete week from inferred home and work locations, individual weekly tour, dwell, and burst parameters, and rank-based exploration and preferential return. Its original task is full-trajectory generation rather than missing-span reconstruction. Following the scikit-mobility reference implementation, we fit a population-level out-of-home rhythm on the GPS training split and retain rank-based exploration and preferential return at the paper's default parameters, preserve the dictated episodes, and restrict simulation to the intervening free intervals. TimeGeo carries no activity channel, so non-home activity labels are sampled from the corpus non-home activity mix and durations from a fitted stay-duration histogram.
- **MIRAGE** [25] is a task-agnostic event-sequence generator combining a GRU history encoder, a log-normal-mixture temporal point process, a cascading activity–location decoder, a neural exploration–return mechanism, and a user-level VAE. It originally generates new trajectories without conditioning on arbitrary retained events. We use the authors' implementation, represent each POI as a (zone, activity) pair, and clamp dictated events at their exact times inside the autoregressive sampler so that MIRAGE generates only within the remaining gaps.

Controlled generation through native infilling.

- **TrajGPT-DH** [27] formulates controlled synthetic trajectory generation as text-style infilling, predicting visit regions and destination-conditioned travel times and durations between retained visits. Its native infilling task is the closest published formulation to ours, but it does not predict activity semantics. We use the authors' implementation and native BLANK/ANS interface, add an activity embedding, prediction head, and joint loss, and train over the dictated-mask range encountered in weekly deployment.
- **BERT4Traj-DH** [13] uses bidirectional masked reconstruction to recover missing locations in a known user's daily CDR or GPS trajectory, conditioned on spatial, temporal, demographic, and anchor information. Our task instead supplies a dictated synthetic week and requires activity as well as location. Because no source code is released, we reimplement the model,

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

replace the daily sequence with 672 fifteen-minute weekly bins, treat dictated events as the observed context, add separate activity and zone heads, and use role identifiers as the demographic proxy and Space2Vec-style multi-scale features of MDS zone coordinates as the coordinate-based location embedding.

- **AttnMove-DH** [28] recovers missing locations in one target day by combining the observed portions of that day with preceding daily trajectories from the same user through intra- and inter-trajectory attention. Because synthetic residents have no longitudinal mobility history, we reimplement AttnMove using the available dictated-week context. Inter-trajectory attention operates across the seven days of the same week at matched times of day, and the output is extended from daily locations to weekly activity–zone tokens. This row evaluates the adapted AttnMove attention architecture under dictated-week inputs, rather than the original multi-day personal-history setting.
- **DeepTimeGeo-DH** [12] reconstructs sparse hourly LBS trajectories using an encoder-only Transformer, mobility-distribution regularizers, exploration–return supervision, and a personalized location-rank vocabulary derived from each observed user's stay points. We use the authors' implementation but replace its multi-week hourly slot array with a 672-bin fifteen-minute week and refit its circadian and rank distributions on our GPS training split. Because a synthetic resident has no observed location history, its per-user rank vocabulary, capped at 45 ranks, is seeded only from the zones of the dictated bins, with the home and work or school anchors appended when unseen; dictated bins form the observed skeleton, decoding follows their deterministic argmax protocol, and an additional activity head provides the semantic output.

Task-generic references.

- **Semi-Markov** models transitions between activities together with per-activity dwell-time distributions and fills the free intervals between dictated events; zones are drawn from a home-distance prior with a per-person revisit urn.
- **LSTM** is a bidirectional recurrent sequence encoder trained from scratch on the common weekly completion examples with separate activity and zone heads.
- **Transformer** replaces recurrent propagation with self-attention while using the same weekly inputs, training examples, and activity and zone heads as the LSTM.

HEARTH achieves the lowest average divergence among the nine evaluated baselines, with Std-6 of 0.0252 compared with 0.0412 for MIRAGE. The advantage is concentrated in spatial structure: HEARTH reduces DailyLoc divergence to 0.0381 from 0.0568 for the nearest baseline, Radius to 0.0164 from 0.0343, and OD-flow divergence to 0.3007 from 0.3329. Temporal and semantic performance is more mixed: MIRAGE is slightly better on Duration (0.0520 versus 0.0553), and Semi-Markov is better on Category (0.0064 versus 0.0099). This pattern is consistent with HEARTH's anchor-conditioned excursion representation contributing most strongly to destination choice, mobility range, and flow structure, while specialized baselines retain narrow advantages in duration or activity-share modeling.

MIRAGE is the closest baseline overall, consistent with its continuous-time generation and exploration–return mechanism, whereas TrajGPT-DH is the closest competitor on OD flow. DeepTimeGeo-DH preserves frequencies over its seeded venue ranks but performs poorly when generating movement beyond those observed ranks, reaching a Std-6 average of 0.2142. Parameter count does not explain the remaining gap: scaling the LSTM from 2.05M to 54.0M parameters reduces validation loss from 6.02 to 0.81 but changes Std-6 from 0.0756 to 0.0783. The comparison therefore points to representation and decoding structure, rather than recurrent-model capacity alone, as the main source of the performance difference. Supplementary Section SVIII reports parameter counts and the complete capacity-control experiment.

7.4. Component ablations

7.4.1. Encoder ablation

Every downstream module consumes the encoder's memory, so we decompose its contribution into architecture and pretraining with two matched retrainings, identical data, steps, seed, and decode. E1 keeps the full encoder architecture and trains from random initialization. E2 reduces the encoder to a plain 6-layer Transformer of the same width, removing the MDS geometry, POI affordances, anchor cross-attention, and MoE, again from random initialization. The four spatial components enter and leave together, so this ablation validates them as a bundle; the inference-time POI ablation of Section 6.2 isolates the place-semantics channel separately. Table 14 reports both against the shipped warm-started encoder (E0).

Architecture contributes −29% and pretraining another −30%. The move from E2 to E1 improves the average from 0.0503 to 0.0358, concentrated exactly where the spatial machinery aims: DailyLoc falls 45% and Radius 51%. The warm start then carries 0.0358 to 0.0252. Substituting the plain encoder and forgoing pretraining together doubles the error of the shipped model, so the encoder is an earned component on both axes. Validation cross-entropy ranked E1 and E2 as equals (2.576 vs. 2.580); the deployment metric separates them by 29%, a reminder that token-level fit and deployment realism are different quantities.

7.4.2. Decoder ablation

The excursion decoder emerged through a staged development campaign. Table 15 traces the decoder ladder with the encoder held fixed and each column accumulating one design change from left to right. The bin-grid variant

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

Table 13
Trajectory-level realism on the deployment task: six-metric JSD suite, full 395K-week decode, 50K-sample evaluation against the held-out GPS test split. Lower is better; best per row in **bold**. Transf. = Transformer, S-Markov = Semi-Markov, B4T = BERT4Traj, AttnM = AttnMove, DTG = DeepTimeGeo; -DH marks task-adapted completion systems with dual activity/zone heads. Subsampling bootstrap intervals on HEARTH (B=20 replicates of 40K samples) give per-metric half-widths of 0.0002–0.0010.

Metric	**HEARTH (ours)**	MIRAGE	TrajGPT-DH	TimeGeo	LSTM	Transf.	B4T-DH	AttnM-DH	S-Markov	DTG-DH
DailyLoc	**0.0381**	0.0779	0.0568	0.1618	0.1896	0.2561	0.2515	0.2663	0.4105	0.3293
Departure	**0.0286**	0.0319	0.0529	0.0328	0.0408	0.0555	0.0610	0.0527	0.0493	0.0468
Duration	0.0553	**0.0520**	0.0912	0.0869	0.0712	0.0791	0.0810	0.0853	0.0559	0.1786
Distance	**0.0030**	0.0292	0.0268	0.0505	0.0113	0.0235	0.0203	0.0268	0.0523	0.1886
Radius	**0.0164**	0.0365	0.0343	0.0447	0.0941	0.1211	0.1208	0.1253	0.1138	0.1975
Category	0.0099	0.0197	0.0438	0.0181	0.0465	0.0246	0.0223	0.0246	**0.0064**	0.3445
Std-6 avg	**0.0252**	0.0412	0.0510	0.0658	0.0756	0.0933	0.0928	0.0968	0.1147	0.2142
OD flow	**0.3007**	0.3812	0.3329	0.5290	0.4253	0.4739	0.4605	0.4803	0.5373	0.6548

Table 14
Encoder ablation, full-decode evaluation under a fixed ablation decode configuration. E0 = shipped encoder (warm start); E1 = full architecture, random initialization; E2 = plain Transformer, random initialization.

Metric	E0	E1	E2
DailyLoc	**0.0381**	0.0510	0.0924
Departure	**0.0286**	0.0606	0.0667
Duration	**0.0553**	0.0614	0.0692
Distance	0.0030	0.0034	**0.0013**
Radius	**0.0164**	0.0311	0.0632
Category	0.0099	**0.0082**	0.0088
Std-6 avg	**0.0252**	0.0358	0.0503

Table 15
Decoder ablation under a fixed comparison configuration, with the encoder held fixed and application-level controls off. Columns accumulate design changes from left to right; **bold** marks the best value per row. The mixture-timing-heads column is a pre-registered negative result retained for the record.

Metric	Bin-grid	Episode AR	+persona	+timing heads	Excursion
DailyLoc	**0.0201**	0.0457	0.0377	0.0353	0.0381
Departure	0.0489	0.0546	0.0416	0.0367	**0.0286**
Duration	0.0911	0.0897	0.0831	0.0832	**0.0553**
Distance	0.1469	0.0168	0.0058	0.0049	**0.0030**
Radius	0.0363	0.0290	0.0221	0.0208	**0.0164**
Category	0.0180	0.0265	0.0174	0.0202	**0.0099**
Std-6 avg	0.0602	0.0437	0.0346	0.0335	**0.0252**

decodes 50K weeks; every episode-based variant decodes the full 395K-week shard.

The decoding geometry and the heads inside it carry the gains. Replacing bin-level mask-predict with an episode-autoregressive decoder that tiles the anchor-bounded gaps improves the average from 0.0602 to 0.0437 and resolves the bin decoder's spatial collapse in one step (Distance 0.1469 to 0.0168). The persona latent improves every cell and brings the average to 0.0346. The timing-heads column is the informative negative result: swapping the categorical duration head for the log-normal mixture while keeping gap tiling leaves Duration essentially unchanged (0.0831 to 0.0832), even as Departure improves. Excursion decoding, sampling from those same mixture heads with home as the ambient state, then reduces Duration to 0.0553 and brings the average from 0.0335 to 0.0252.

7.5. Granular distributional fidelity

Averages summarize performance, but distributions reveal where the remaining differences arise. Figure 12 compares eight distributions from the shipped decode with the held-out GPS test split. The temporal results reflect the distinction between dictated and generated behavior. Departure times retain the morning and evening peaks established by the mandatory schedules, although the generated distribution places slightly more mass at these peaks and less around midday (JSD 0.0286). The dictated portion provides the main timing structure, while HEARTH generates the remaining variation. Episode durations reproduce both the short activity mode near one hour and the workday mode near eight to nine hours (0.0179). This result supports the use of episode decoding with mixture duration heads: in the decoder ablation of Table 15, bin-level decoding scores 0.0911 on the Std-6 Duration cell against 0.0553 for the shipped decoder. Category shares also remain close overall (0.0099), with lower Work and higher Recreation as the main visible differences. Mean temporal occurrence is more accurate on weekdays than on weekends (0.0245 versus 0.0495). This result is consistent with the cross-level analysis in Section 6.1: weekdays contain more dictated anchors, whereas weekend behavior is generated more freely.

The spatial distributions also agree closely with the reference, with a mild tendency toward more local travel. Zone visit frequencies follow the reference across all 512 zones, including the low frequency tail (0.0250). The number of unique zones visited per day retains the dominant two-zone mode (0.0193), although the generated trajectories contain slightly more one-zone days and fewer days with several zones. Zone dwell times reproduce the short stay mode and the secondary mode near eight hours (0.0248). Daily travel distance follows the reference decay into the long

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

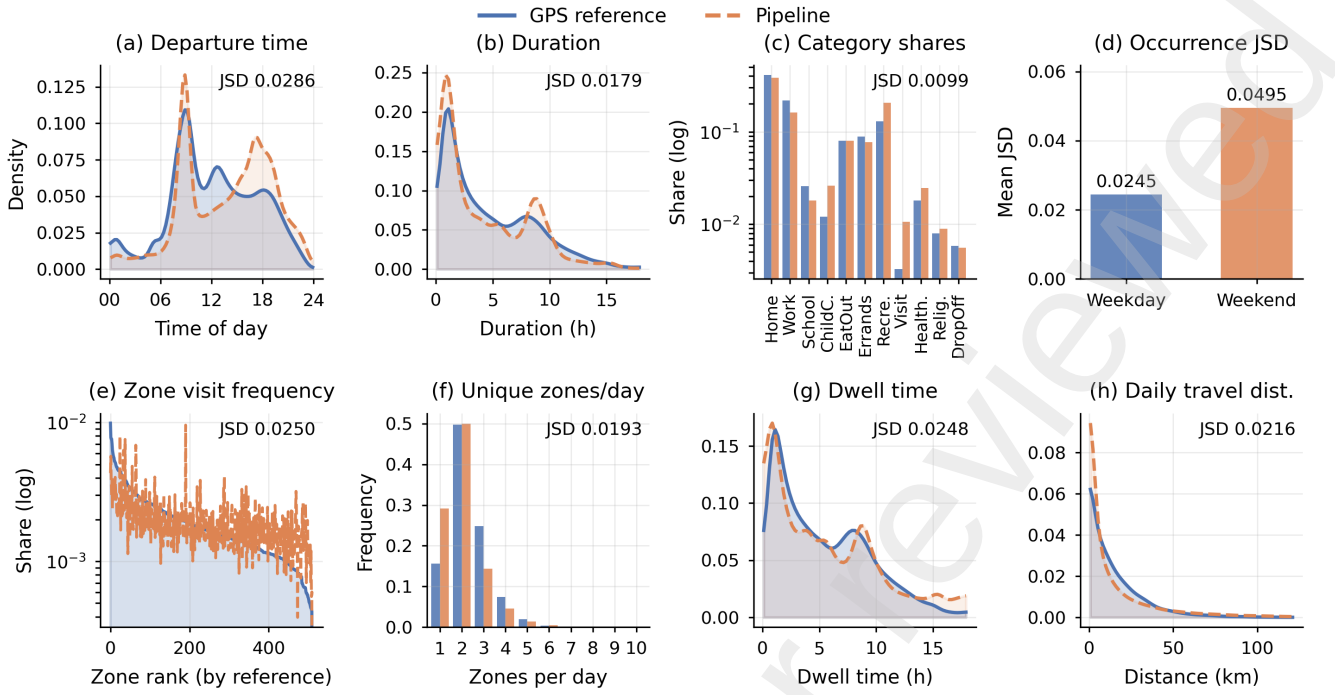

Figure 12: Granular fidelity of the shipped decode against the held-out GPS test split in JSD, the same reference and evaluation protocol as Table 13. Top row, temporal: (a) episode departure time, (b) episode duration, (c) activity category shares, and (d) mean temporal-occurrence JSD by weekday and weekend. Bottom row, spatial: (e) zone visit frequencies across all 512 zones, ranked by reference share, (f) unique zones per day, (g) zone dwell time, and (h) daily travel distance. Panel (b) pools episodes across activities, whereas the Std-6 Duration cell averages the eleven activity-specific divergences. Panel (f) pools all person-days, whereas the Std-6 DailyLoc cell averages the seven day-of-week-specific divergences.

tail, with a small excess of very short travel days (0.0216). Together, the additional one-zone days and short travel days suggest that the model generates slightly fewer long chained tours. Tour chaining is therefore a natural target for further improvement.

Table 16 separates two aspects of the activity fit. The *weekly profile* measures when an activity occurs. It is the population occupancy curve over the week's 672 15-minute bins, normalized over the week and compared with the reference by JSD. The *participation rate* measures how many people perform the activity at least once on each day. We report the RMSE between generated and reference participation rates across the seven days.

The two metrics reveal different error patterns across activities. Healthcare has the lowest profile divergence at 0.0235, while Religious has the lowest participation error at 0.0163. Both activities therefore agree closely with the reference in timing and participation. Recreation has a moderate profile divergence of 0.0389 but the largest participation error at 0.243. Together with its higher activity share in panel (c), this indicates that the model places Recreation at plausible times but assigns it to too many people. Visit and DropOff show the reverse pattern. Their participation errors are smaller at 0.074 and 0.021, but their profile divergences are the largest at 0.1086 and 0.1185. The model therefore estimates participation more accurately than placement within the week for these sparse activities. ChildCare, EatOut, and Errands fall between these cases.

Table 16
Per-discretionary-activity fit of the shipped decode against the held-out GPS test split: JSD of the 672-bin weekly occupancy profile and RMSE of the seven per-day participation rates (fraction of persons performing the activity at least once per day).

Activity	Profile JSD	Participation RMSE
ChildCare	0.0600	0.061
EatOut	0.0472	0.099
Errands	0.0272	0.090
Recreation	0.0389	0.243
Visit	0.1086	0.074
Healthcare	0.0235	0.0382
Religious	0.0348	0.0163
DropOff	0.1185	0.021

Figures 13 and 14 measure activity–location-coupled granular fidelity. Supplementary Section SIX provides the expanded distance and activity-conditioned venue-entropy panels. Trip distances split by day type stay within 0.0026 (weekday) and 0.0032 (weekend) of the reference, and home-to-activity distances for the four everyday discretionary classes land between 0.0013 and 0.0024 (Fig. 13), the distributions the anchor conditioning

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

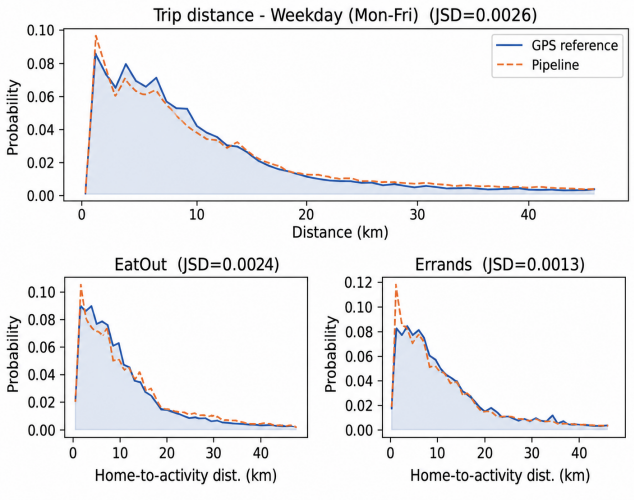

Figure 13: Distance structure of the shipped decode. Top: weekday trip-distance distribution (JSD 0.0026; weekend 0.0032). Bottom: home-to-activity distance for EatOut (0.0024) and Errands (0.0013); Recreation (0.0013) and Healthcare (0.0020) appear in the supplementary material.

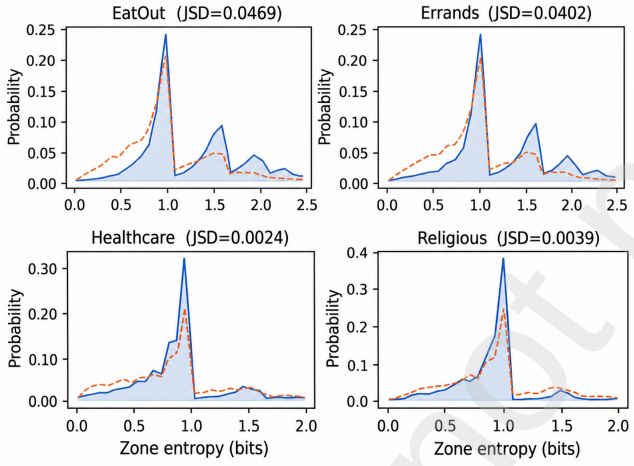

Figure 14: Activity-conditioned per-person zone entropy, shipped decode against the GPS test split, for the two loosest classes (EatOut 0.0469, Errands 0.0402) and two of the tightest (Healthcare 0.0024, Religious 0.0039). All eight discretionary classes appear in the supplementary material (Visit 0.0029, DropOff 0.0021, ChildCare 0.0183, Recreation 0.0083).

and excursion geometry exist to produce. Habits survive generation too. Fig. 14 conditions per-person zone entropy on activity, asking whether synthetic residents concentrate each activity at habitual venues the way real residents do. The concentration structure is reproduced across all eight discretionary classes, tightest for Visit, Healthcare, Religious, and DropOff (0.0021–0.0039); for EatOut (0.0469) and Errands (0.0402) the model spreads modest extra mass into higher-entropy, multi-venue patterns.

Table 17
Matched-resolution HEARTH retraining at 512 and 1,024 zones. Both models reconstruct held-out GPS weeks from the same mandatory-anchor view and are scored at their native resolution. Lower is better.

Metric	512 zones	1,024 zones	Interpretation
DailyLoc	0.0262	0.0424	granularity-sensitive
Departure	0.0068	0.0094	nearly flat
Duration	0.0281	0.0323	mildly higher
Distance	0.0036	0.0023	improved
Radius	0.0048	0.0045	slightly improved
Category	0.0084	0.0092	nearly flat
Std-6 average	**0.0130**	**0.0167**	main evaluation suite

7.6. Choice of spatial resolution

The pipeline operates on 512 zones, corresponding to approximately 4 km^2 per zone and a scale comparable to a traffic-analysis zone. Finite-sample evaluation noise increases with spatial resolution: the zone-frequency and OD-flow floors rise from 0.0132 and 0.1773 at 512 zones to 0.0245 and 0.2454 at 1,024 zones. At 512 zones, these floors remain below the corresponding deployed model–reference divergences of 0.0250 and 0.3007, while the training corpus retains a median of 762 observed transitions per zone. This resolution therefore preserves meaningful spatial variation without making the deployment evaluation predominantly sampling-noise limited. Supplementary Section SVII reports the full sampling-noise sweep from 128 to 4,096 zones.

We test the effect of a finer partition through matched retraining. The GPS corpus is rebuilt at 1,024 zones by subdividing each 512-zone cell, and every zone-derived artifact—centroids, the distance matrix, MDS coordinates, and POI affordances—is regenerated at the finer resolution. HEARTH is then trained from scratch with the same architecture and optimization recipe. Both models reconstruct held-out GPS person-weeks from the same mandatory-anchor view, which retains each week's Work and School episodes together with its home and work or school anchors and masks the remaining content. Each model is scored at its native spatial resolution. Because this experiment uses observed rather than generated upstream inputs, it is oracle-style rather than a deployed generated-input comparison. It is nevertheless a separate matched-resolution retraining and evaluation run, so its 512-zone Std-6 of 0.0130 is not the 0.0146 GPS-dictated oracle result in Table 8.

Finer spatial resolution does not degrade all metrics uniformly. The increase in Std-6 from 0.0130 at 512 zones to 0.0167 at 1,024 zones is concentrated in DailyLoc, which rises from 0.0262 to 0.0424; Distance and Radius remain stable or improve, and the temporal and category metrics change by less than 0.005. Zone-frequency JSD also grows by 1.9×, closely matching the 1.9× increase in its sampling-noise floor. The matched retraining therefore indicates that

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

the apparent resolution penalty arises primarily from estimating person-day location diversity and sparse zone frequencies, rather than from a general loss of temporal or movement fidelity.

Across the five research questions, a common pattern emerges: each source-aligned component primarily preserves the relationship it is designed to carry. The persona prior supplies demographic separation, while the household-schedule latent supplies demographic and household dependence that aggregate marginal metrics do not reveal (RQ1–RQ3). The reference-shift and source-withholding experiments show that the principal conclusions are stable, while also distinguishing inference-time source dependence from training-time contribution (RQ4). Module-level benchmarks establish competitive performance on each level's defined task, and the GPS-dictated diagnostic locates a substantial share of the remaining end-to-end error in the generated upstream inputs (RQ5).

8. Discussion

The results establish PopTraj's hierarchical factorization at metropolitan scale while defining its empirical scope. Validation currently covers one metropolitan area, with local movement and spatial behavior learned from its GPS panel; census and survey evidence separately govern population structure and group-level time allocation. GPS activity labels follow the consistent weakly supervised pipeline of Section 5.1, and cross-region evaluation with additional mobility sources is a natural next step. The remaining trajectory errors concentrate in activity participation, particularly excess Recreation and reduced weekday Work, rather than in the overall temporal and spatial structure, identifying a focused direction for refinement.

The two-token anchor memory is asymmetric for residents who both work and study. In a 60K-record deployment audit, 8.29% of residents carried both Work and School episodes. Their primary mandatory-anchor token stores Work, while the School zone remains visible through the fixed dictated School bins. HEARTH therefore preserves their School visits but does not expose School as a global anchor when locating discretionary excursions. In addition, 6.54% of records carry a Work anchor without a Work episode in the sampled week, so the prefix represents habitual spatial context rather than a claim that the activity occurs during that week. Extending the prefix to expose both mandatory anchors is a direction for future work.

PopTraj is designed for population-level simulation rather than prediction of identifiable individuals: each trajectory is a plausible synthetic draw whose fidelity is evaluated at aggregate and demographic-group levels. Demographic conditioning should consequently be interpreted and monitored at those levels, especially for sparsely represented strata. The teacher corpus also exhibits a localized discrepancy in the activity composition of parent–child events, although total joint time remains aligned with the survey. The language model is absent from deployment, but its filtered proposals enter the distilled schedule prior. This discrepancy should therefore be treated as a measured bias in the Level-II training target whose downstream effect is assessed through the schedule and end-to-end evaluations.

9. Conclusion

We presented PopTraj, a hierarchical multi-source generation framework that turns census constraints, time-use targets, geographic data, and anonymous GPS observations into household-coordinated weekly activity–location trajectories for synthetic urban populations. Level I constructs household populations and spatial anchors, Level II generates coordinated weekly schedules, and Level III uses HEARTH to complete those schedules through home-anchored excursions and demographic personas. In the Tokyo deployment, HEARTH reaches 0.0252 mean divergence versus 0.0412 for the strongest of nine baselines; the deployed schedule pipeline reduces coordination error by more than 16× relative to six learned baselines; and the persona prior halves demographic divergence with little change in aggregate realism. Household venue alignment reaches 0.994 shared-event co-presence, and error-propagation analysis locates the largest remaining error upstream. The framework decodes hundreds of thousands of weeks in hours without a language model at inference. Future work will focus on the measured participation errors, persona coverage for weeks with few mandatory anchors, and validation across additional regions.

CRediT authorship contribution statement

Haoxuan Ma: Conceptualization, Methodology, Software, Investigation, Writing – original draft. **Yifan Liu:** Methodology, Software, Investigation, Writing – original draft. **Bo Yang:** Data curation, Software. **Yiwen Zhang:** Software, Validation. **Chris Stanford:** Methodology, Writing – review & editing. **Eli Fox:** Methodology, Writing – review & editing. **Xishun Liao:** Methodology, Writing – review & editing. **Jiaqi Ma:** Supervision, Funding acquisition, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0033. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

Data availability

The census marginals (e-Stat) and time-use survey statistics (JTUS-2016) are public. The anonymized GPS corpus was obtained under a data agreement and cannot be redistributed; the synthetic population, generated schedules, and completed trajectories contain no real individual and can be shared. Code for the population synthesizer, schedule models, and HEARTH will be released upon publication.

References

[1] A. Horni, K. Nagel, K. W. Axhausen (Eds.), The Multi-Agent Transport Simulation MATSim, Ubiquity Press, London, 2016. doi:10.5334/baw.

[2] M. J. Roorda, E. J. Miller, K. M. N. Habib, Validation of TASHA: A 24-h activity scheduling microsimulation model, Transportation Research Part A: Policy and Practice 42 (2) (2008) 360–375. doi:10.1016/j.tra.2007.10.004.

[3] S. M. Reia, H. F. de Arruda, S. Ruan, T. Anderson, H. Kavak, D. Pfoser, Towards universal urban patterns-of-life simulation (2026). arXiv:2601.22099, doi:10.48550/arXiv.2601.22099.

[4] N. E. Klepeis, W. C. Nelson, W. R. Ott, J. P. Robinson, A. M. Tsang, P. Switzer, J. V. Behar, S. C. Hern, W. H. Engelmann, The national human activity pattern survey (NHAPS): A resource for assessing exposure to environmental pollutants, Journal of Exposure Analysis and Environmental Epidemiology 11 (3) (2001) 231–252. doi:10.1038/sj.jea.7500165.

[5] C. R. Bhat, J. Y. Guo, S. Srinivasan, A. Sivakumar, Comprehensive econometric microsimulator for daily activity-travel patterns, Transportation Research Record 1894 (1) (2004) 57–66. doi:10.3141/1894-07.

[6] V. Kulkarni, N. Tagasovska, T. Vatter, B. Garbinato, Generative models for simulating mobility trajectories, workshop on Modeling and Decision-Making in the Spatiotemporal Domain, NeurIPS 2018 (2018). arXiv:1811.12801, doi:10.48550/arXiv.1811.12801.

[7] J. Feng, Z. Yang, F. Xu, H. Yu, M. Wang, Y. Li, Learning to simulate human mobility, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3426–3433. doi:10.1145/3394486.3412862.

[8] S. Choi, J. Kim, H. Yeo, TrajGAIL: Generating urban vehicle trajectories using generative adversarial imitation learning, Transportation Research Part C: Emerging Technologies 128 (2021) 103091. doi:10.1016/j.trc.2021.103091.

[9] Y. Zhu, Y. Ye, S. Zhang, X. Zhao, J. Yu, DiffTraj: Generating GPS trajectory with diffusion probabilistic model, in: Advances in Neural Information Processing Systems, Vol. 36, 2023, pp. 65168–65188. doi:10.52202/075280-2843.

[10] J. Wang, R. Jiang, C. Yang, Z. Wu, M. Onizuka, R. Shibasaki, N. Koshizuka, C. Xiao, Large language models as urban residents: An LLM agent framework for personal mobility generation, in: Advances in Neural Information Processing Systems, Vol. 37, 2024, pp. 124547–124574. doi:10.52202/079017-3957.

[11] H. Ma, X. Liao, Y. Liu, Q. Jiang, C. Stanford, S. Cao, J. Ma, Learning universal human mobility patterns with a foundation model for cross-domain data fusion, Transportation Research Part C: Emerging Technologies 180 (2025) 105311. doi:10.1016/j.trc.2025.105311.

[12] S. Cao, J. Wu, A. Kasliwal, B. Chen, G. Perona, M. C. González, Deeptimegeo: Trajectory reconstruction from sparse data with transformer, IEEE Transactions on Intelligent Transportation Systems 27 (5) (2026) 5586–5599. doi:10.1109/TITS.2026.3657275.

[13] H. Yang, A. Yao, C. Whalen, G. Mai, BERT4Traj: Transformer based trajectory reconstruction for sparse mobility data (2025). arXiv:2507.03062, doi:10.48550/arXiv.2507.03062.

[14] W. Luo, C. Xiong, Recovering individual-level activity sequences from location-based service data using a novel transformer-based model (2025). arXiv:2508.02734, doi:10.48550/arXiv.2508.02734.

[15] H. Orchi, A. B. Diallo, H. Elbiaze, E. Sabir, M. Sadik, A contemporary survey on multisource information fusion for smart sustainable cities: Emerging trends and persistent challenges, Information Fusion 114 (2025) 102667. doi:10.1016/j.inffus.2024.102667.

[16] Y. Yan, S. Cui, J. Liu, Y. Zhao, B. Zhou, Y.-H. Kuo, Multimodal fusion for large-scale traffic prediction with heterogeneous retentive networks, Information Fusion 114 (2025) 102695. doi:10.1016/j.inffus.2024.102695.

[17] B. Matet, E. Côme, A. Furno, S. Hörl, L. Oukhellou, N.-E. El Faouzi, Improving the generation of synthetic travel demand using origin–destination matrices from mobile phone data, Transportation 53 (3) (2026) 1107–1139. doi:10.1007/s11116-024-10524-2.

[18] K. D. Vo, E.-J. Kim, P. Bansal, A novel data fusion method to leverage passively-collected mobility data in generating spatially-heterogeneous synthetic population, Transportation Research Part B: Methodological 191 (2025) 103128. doi:10.1016/j.trb.2024.103128.

[19] Ç. Tozluoğlu, S. Dhamal, S. Yeh, F. Sprei, Y. Liao, M. Marathe, C. L. Barrett, D. Dubhashi, A synthetic population of Sweden: datasets of agents, households, and activity-travel patterns, Data in Brief 48 (2023) 109209. doi:10.1016/j.dib.2023.109209.

[20] J. de Mooij, T. Sonnenschein, M. Pellegrino, M. Dastani, D. Ettema, B. Logan, J. A. Verstegen, Gensynthpop: generating a spatially explicit synthetic population of individuals and households from aggregated data, Autonomous Agents and Multi-Agent Systems 38 (2) (2024) 48. doi:10.1007/s10458-024-09680-7.

[21] M. Bradley, P. Vovsha, A model for joint choice of daily activity pattern types of household members, Transportation 32 (5) (2005) 545–571. doi:10.1007/s11116-005-5761-0.

[22] F. Shone, T. Hillel, Black box behavioural modelling: Predicting human activity schedules with a deep conditional generative approach (2025). arXiv:2512.04223, doi:10.48550/arXiv.2512.04223.

[23] Y. Yuan, J. Ding, H. Wang, D. Jin, Y. Li, Activity trajectory generation via modeling spatiotemporal dynamics, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 4752–4762. doi:10.1145/3534678.3542671.

[24] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, M. C. González, The TimeGeo modeling framework for urban mobility without travel surveys, Proceedings of the National Academy of Sciences 113 (37) (2016) E5370–E5378. doi:10.1073/pnas.1524261113.

[25] B. Deng, X. Jing, T. Yang, B. Qu, D. Yang, P. 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, ACM, 2025, pp. 201–212. doi:10.1145/3690624.3709180.

[26] Y. Luo, S. Zhang, K. Liu, Y. Xu, L. Yin, Traveller: Travel-pattern aware trajectory generation via autoregressive diffusion models, Information Fusion 127 (2026) 103766. doi:10.1016/j.inffus.2025.103766.

[27] S.-L. Hsu, E. Tung, J. Krumm, C. Shahabi, K. Shafique, TrajGPT: Controlled synthetic trajectory generation using a multitask transformer-based spatiotemporal model, in: Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, ACM, 2024, pp. 362–371. doi:10.1145/3678717.3691303.

[28] T. Xia, Y. Qi, J. Feng, F. Xu, F. Sun, D. Guo, Y. Li, AttnMove: History enhanced trajectory recovery via attentional network, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 4494–4502. doi:10.1609/aaai.v35i5.16577.

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

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085

[30] X. Liao, Y. Liu, C. Kuai, H. Ma, B. Y. He, S. Cao, C. Stanford, J. Ma, Reconstructing human mobility pattern: A semi-supervised approach for cross-dataset transfer learning (2024). arXiv:2410.03788, doi:10.48550/arXiv.2410.03788.

[31] A. G. Wilson, A statistical theory of spatial distribution models, Transportation Research 1 (3) (1967) 253–269. doi:10.1016/0041-1647(67)90035-4.

[32] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, β-VAE: Learning basic visual concepts with a constrained variational framework, international Conference on Learning Representations (ICLR) (2017).
URL https://openreview.net/forum?id=Sy2fzU9gl

[33] W. S. Torgerson, Multidimensional scaling: I. theory and method, Psychometrika 17 (4) (1952) 401–419. doi:10.1007/BF02288916.

[34] OpenStreetMap contributors, OpenStreetMap, https://www.openstreetmap.org, data available under the Open Database License (ODbL) (2024).

[35] Y. Liu, C. Kuai, X. Liao, H. Ma, B. Y. He, J. Ma, Semantic trajectory data mining with LLM-informed POI classification, in: Proceedings of the IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), 2024, pp. 207–213. doi:10.1109/ITSC58415.2024.10920138.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, 2017, pp. 5998–6008.
URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[37] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, Y. Liu, RoFormer: Enhanced transformer with rotary position embedding, Neurocomputing 568 (2024) 127063. doi:10.1016/j.neucom.2023.127063.

[38] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, international Conference on Learning Representations (ICLR) (2017).
URL https://openreview.net/forum?id=B1ckMDqlg

[39] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[40] A. van den Oord, O. Vinyals, K. Kavukcuoglu, Neural discrete representation learning, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 6306–6315.
URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html

[41] B. Yang, H. Ma, Y. Liu, Z. Zhang, C. Stanford, M. Sun, J. Ma, Uncertainty-aware trip purpose inference from GPS trajectories via POI semantic zones and pareto calibration (2026). arXiv:2605.01257, doi:10.48550/arXiv.2605.01257.

[42] Statistics Bureau of Japan, e-Stat: Portal site of official statistics of japan, https://www.e-stat.go.jp/en, ministry of Internal Affairs and Communications (2024).

[43] Statistics Bureau of Japan, 2016 survey on time use and leisure activities, https://www.stat.go.jp/english/data/shakai/2016/gaiyo.html, ministry of Internal Affairs and Communications (2017).

[44] W. E. Deming, F. F. Stephan, On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, The Annals of Mathematical Statistics 11 (4) (1940) 427–444. doi:10.1214/aoms/1177731829.

[45] J. Barthelemy, P. L. Toint, Synthetic population generation without a sample, Transportation Science 47 (2) (2013) 266–279. doi:10.1287/trsc.1120.0408.

[46] F. Simini, M. C. González, A. Maritan, A.-L. Barabási, A universal model for mobility and migration patterns, Nature 484 (7392) (2012) 96–100. doi:10.1038/nature10856.

[47] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein GANs, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 5769–5779.
URL https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7484085