When Learned Compression Loses Identity: Stage-Resolved Diagnostics for Exact-Entity Conditional Generation

Anonymous authors
Paper under double-blind review

Abstract

Learned compression is effective when task-valid variation can be discarded without identity loss, but can become a misaligned inductive bias when correctness depends on exact, non-substitutable categorical identity. Existing evidence does not determine where exact identity is lost, while comparisons that jointly change representation and corruption cannot attribute failure to either factor. We study personalized multi-day mobility generation as an identity-sensitive testbed, where recent user history conditions future visits over a known location vocabulary, yet models may produce structurally plausible sequences with the wrong locations for that user. An ordered representation–corruption intervention with stage-specific probes localizes identity loss and tests alternative explanations. The diagnosis reveals two distinct mechanisms: vector quantization aliases exact locations before diffusion, whereas continuous autoencoding preserves near-perfect clean reconstruction yet loses prefix-specific identity access during Gaussian latent denoising. Guided by this diagnosis, NADI instantiates the native-mask endpoint through native-space absorbing diffusion and achieves the highest composite objective across all four datasets in the primary evaluation. More broadly, reconstruction fidelity alone does not establish the suitability of a generative representation: for exact, non-substitutable entities, representation and corruption should be judged by whether condition-supplied identity distinctions remain accessible throughout generation.

1 Introduction

Learned compression underlies many successful generative models, including vector-quantized autoencoders and latent diffusion (van den Oord et al., 2017; Rombach et al., 2022). Its effectiveness reflects a useful inductive bias: representations may discard fine-grained variation that is not fixed by the condition while preserving what is sufficient for a valid output. In text-conditioned image generation, for example, viewpoint, texture, or lighting may vary without changing the prompt-specified content. This tolerance disappears when correctness depends on the exact categorical identity of the generated entity, because even a plausible alternative class constitutes an identity error (Figure 1). We therefore ask: *when can learned compression become the wrong inductive bias for preserving exact categorical identity in conditional generation?* Here and throughout, identity refers to the exact task-defined output class, not to the identification of a person.

This question defines a failure regime in which aggregate structural fidelity can conceal condition-specific identity loss. Prior work establishes the ingredients for several plausible mechanisms: reconstruction fidelity need not predict latent diffusability (Yao et al., 2025; Skorokhodov et al., 2025); quantization can alias distinct items under a shared semantic identifier (Rajput et al., 2023; Zhang et al., 2026); and representation geometry, semantic information, and condition influence can change across denoising (Zhong et al., 2026; Kadkhodaie et al., 2024; Han et al., 2025; Sclocchi et al., 2025; Görgün et al., 2025). Latent-free generation further shows that conventional learned latent interfaces are not necessary (Hoogeboom et al., 2023; Gu et al., 2024; Chen et al., 2025; Li et al., 2024). What none of this establishes is which mechanism is responsible for a particular failure. Two stage-specific mechanisms are especially plausible: *pre-diffusion aliasing* and *denoising-stage loss of identity access from the supplied condition*. Broader alternatives include *insufficient*

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

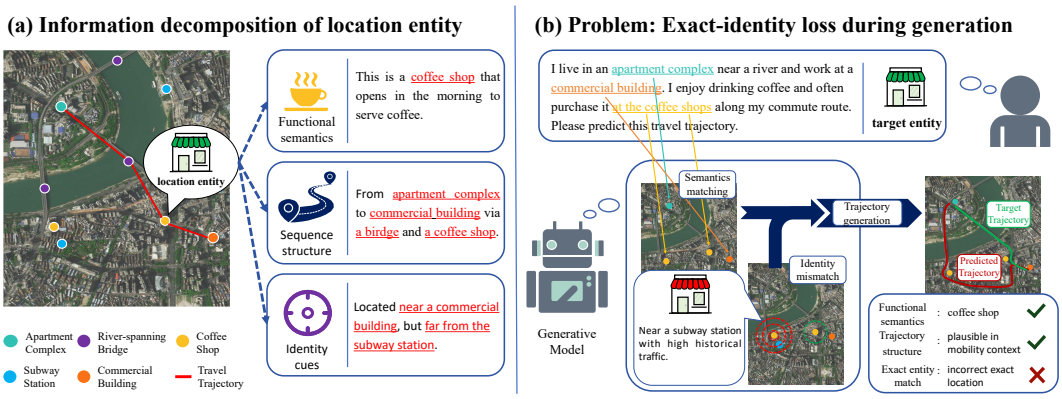

Figure 1: **Exact-entity conditional generation in personalized mobility.** (a) A location entity combines functional semantics, sequence structure, and identity cues. (b) The text schematically verbalizes potential mobility context implied by a user's observed trajectory history; it is not a language prompt or model input. A generated trajectory may remain semantically valid and structurally plausible while selecting the wrong exact location.

representational capacity and *weak conditioning use*. The two stage-specific mechanisms demand different remedies: pre-diffusion aliasing requires changing the representation, whereas denoising-stage loss cannot be repaired by a better encoder. Yet standard evidence leaves both the stage and the cause ambiguous: reconstruction and endpoint quality do not localize where identity is lost, while latent-versus-native comparisons cannot attribute the gap because they change representation and corruption simultaneously. Resolving both ambiguities requires a controlled, stage-resolved design.

We study this problem through personalized multi-day mobility generation, an identity-sensitive testbed in which strong population-level regularities can bias generative models toward structurally plausible yet identity-incorrect futures (González et al., 2008; Song et al., 2010). Given a user's recent visit history, the task generates a future sequence over a known location vocabulary, thereby separating structural fidelity from exact categorical identity. On this testbed, we develop a controlled representation–corruption framework that uses ordered interventions and stage-specific probes to localize identity loss and test alternative explanations. The analysis reveals distinct failure modes across the two learned-latent pathways: vector quantization loses exact identity before diffusion, whereas continuous autoencoding loses condition-specific identity access during denoising. Subsequent controls separate these mechanisms from insufficient representational capacity and generic conditioning failure.

This diagnosis motivates NADI (*Native-space Absorbing Diffusion for Identity-preserving Mobility Generation*), which instantiates the native-mask endpoint through absorbing-state diffusion over native visit tokens (Sahoo et al., 2024; Shi et al., 2024). Its consistent gains across four heterogeneous datasets provide an end-to-end validation of the diagnosis. More broadly, reconstruction fidelity alone does not establish the suitability of a generative representation: when output classes are non-substitutable, representation and corruption must be judged by whether condition-specific categorical identity remains accessible throughout generation.

Contributions.

- We formulate exact-entity conditional generation as a regime in which structural fidelity can conceal categorical identity errors, and introduce personalized multi-day mobility generation as a controlled, identity-sensitive testbed.
- We develop a controlled representation–corruption framework that combines an ordered intervention with stage-specific probes, so that the stage at which identity is lost can be localized and its responsible source isolated.
- Guided by the resulting diagnosis, we propose NADI, a native-space absorbing diffusion model without learned compression, and validate it across four heterogeneous datasets.

Key findings. The two learned-latent pathways fail for different reasons that output-level evidence conflates. Vector quantization destroys identity *before* diffusion: its codes preserve shared temporal regularities almost losslessly while aliasing the locations that carry user identity. By contrast, the continuous autoencoder preserves the scored identity under clean reconstruction, yet that identity largely disappears in generation. Widening the latent until no dimensional reduction remains does not recover it, ruling out insufficient representational capacity as the primary explanation. A user-history swap instead shows that the generated identity becomes largely insensitive to the supplied condition, localizing the failure to condition-specific identity access during denoising.

2 Problem Formulation

Identity-sensitive output space. Each location token $\ell \in \mathcal{V}$ denotes an exact, non-substitutable output class. Exactness concerns the output ontology rather than deterministic reproduction of the realized future: multiple futures may be behaviorally plausible for the same user, but replacing one location class with another constitutes an identity error, even when the substitute is spatially close or plausible at the population level. We therefore evaluate aggregate mobility realism and user-specific location recovery separately.

Structured visit sequences. A visit is represented as

$$v_i = (y_i, w_i), \qquad y_i = (\ell_i, \tau_i, b_i) \in \mathcal{V} \times \mathcal{T} \times \{0, 1\}, \qquad w_i \in \mathcal{W} = \{1, \ldots, 7\},$$

where ℓ_i is the location, τ_i the arrival-time bin, $b_i = 1$ marks the end of a behavioral day, and w_i provides day-of-week context. Behavioral days follow the 04:00 boundary used in prior mobility modeling (Mo et al., 2021); the boundary convention and modality-specific visit construction are detailed in Appendix B.

Personalized multi-day generation. Each instance consists of K preceding behavioral days and the user's next H behavioral days. Let $T_u^{(K)}$ denote the observed prefix and

$$y_u^\star = \big((\ell_i^\star, \tau_i^\star, b_i^\star)\big)_{i=1}^{N}$$

the reference future, where N is the number of visits observed within the H-day reference window. Let $w_{1:N}$ denote the corresponding slot-aligned day-of-week labels. All methods model

$$p_\theta \left(y_{1:N} \mid T_u^{(K)}, N, w_{1:N} \right)$$

and generate

$$\widehat{y}_{1:N} = \big((\widehat{\ell}_i, \widehat{\tau}_i, \widehat{b}_i)\big)_{i=1}^{N}.$$

The reference length N and day-of-week scaffold $w_{1:N}$ are supplied identically to all methods and remain fixed during generation. This isolates visit content, temporal attributes, and day organization while enabling paired representation–corruption interventions over identical output slots. The task therefore evaluates reference-length-conditioned generation rather than autonomous prediction of future visit counts.

The generated boundaries partition the fixed N visits into behavioral days:

$$\widehat{d}(i) = 1 + \sum_{j<i} \widehat{b}_j.$$

A boundary closes the current behavioral day but does not terminate generation. Consequently, total visit volume is conditioned on, whereas visit allocation and organization across days remain generated. Because H specifies only the reference window, the generated partition need not contain exactly H behavioral days.

3 Controlled Diagnostic Framework

We diagnose learned-compression failure by separating two coupled design choices: the representation in which entity distinctions are expressed and the corruption process through which uncertainty is introduced during denoising.

Controlled representation–corruption intervention. Prior work has separately shown that downstream diffusion is sensitive to the geometry and diffusability of learned representations (Skorokhodov et al., 2025; Zhong et al., 2026), and that the corruption kernel materially shapes denoising in discrete state spaces (Austin et al., 2021; Pynadath et al., 2025). A direct comparison between a latent-Gaussian model and a native-mask model changes both factors simultaneously and therefore cannot attribute any performance difference to either representation or corruption. We address this confounding with a deliberately minimal native-Gaussian bridge:

$$\text{CONT+AE} \longrightarrow \text{GAUSS} \longrightarrow \text{MASK}. \tag{1}$$

All pathways follow the common task formulation in Section 2 and share the conditioning protocol and denoiser backbone where compatible. CONT+AE performs Gaussian diffusion in a learned autoencoder latent; GAUSS removes the encoder–decoder while retaining Gaussian corruption over native-token embeddings; and MASK keeps the native representation fixed but replaces Gaussian noise with absorbing masking. Thus, the first contrast in Equation (1) targets the effect of removing learned compression, whereas the second targets the effect of changing corruption.

To align metrics with opposite optimization directions, define

$$s_M = \begin{cases} +1, & \text{if higher values of } M \text{ are better,} \\ -1, & \text{if lower values of } M \text{ are better.} \end{cases} \tag{2}$$

The direction-aligned changes associated with the two sequential steps in Equation (1) are

$$\delta^M_{\text{rep}} = s_M(M_{\text{GAUSS}} - M_{\text{CONT+AE}}), \qquad \delta^M_{\text{mask}} = s_M(M_{\text{MASK}} - M_{\text{GAUSS}}). \tag{3}$$

Under this convention, positive values indicate improvement in M, whereas negative values indicate degradation. The direction-aligned change across the full intervention path is

$$\delta^M_{\text{path}} = \delta^M_{\text{rep}} + \delta^M_{\text{mask}} = s_M(M_{\text{MASK}} - M_{\text{CONT+AE}}). \tag{4}$$

When $\delta^M_{\text{path}} > 0$, we express each sequential change as a signed percentage of the full-path improvement:

$$\rho^M_{\text{rep}} = 100\frac{\delta^M_{\text{rep}}}{\delta^M_{\text{path}}}, \qquad \rho^M_{\text{mask}} = 100\frac{\delta^M_{\text{mask}}}{\delta^M_{\text{path}}}. \tag{5}$$

The two percentages sum to 100%. A negative percentage indicates degradation at that step, whereas a percentage above 100% indicates an improvement large enough to offset degradation at the other step. Because the ladder is sequential rather than factorial, Equations (3)–(5) define an order-dependent diagnostic decomposition rather than order-invariant causal effects. VQ compression and the coordinate-based pathway (DiffTraj) are evaluated separately as orthogonal representation controls, since their representation-specific mechanisms do not form nested single-factor interventions.

Compression-stage audits. To explain the representation-level gap identified by the controlled intervention, we next test whether exact identity is already lost before diffusion. For the continuous autoencoder, we conduct a *clean reconstruction audit*: diffusion is disabled, and exact location reconstruction accuracy is measured after an uncorrupted encode–decode round trip, both overall and within frequency-tail subsets. For the VQ pathway, we measure visit-weighted code purity and conditional collision rates to determine whether a single quantized code aliases multiple locations. Latent-width and codebook-size sweeps provide pathway-specific controls for insufficient representational capacity. The main findings are reported in Section 4.3, with formal definitions and detailed results provided in Appendix E.2.

Prefix-specific identity access during denoising. The clean reconstruction audit tests whether exact location identity survives an uncorrupted encode–decode pathway, but not whether it remains accessible from the supplied user condition once the generative state is corrupted. Let $y^\star_u$ denote the reference future of target user u. For normalized corruption level $r \in (0, 1]$, its pathway-specific corrupted state is

$$\widetilde{s}_{\mathcal{P},r} \sim q_{\mathcal{P},r}(\cdot \mid y^\star_u). \tag{6}$$

Here, $r = 1$ is the maximally corrupted generation start, and smaller values correspond to progressively cleaner states. We exclude $r = 0$, because the uncorrupted state already supplies the target future representation and therefore makes prefix-specific attribution uninformative.

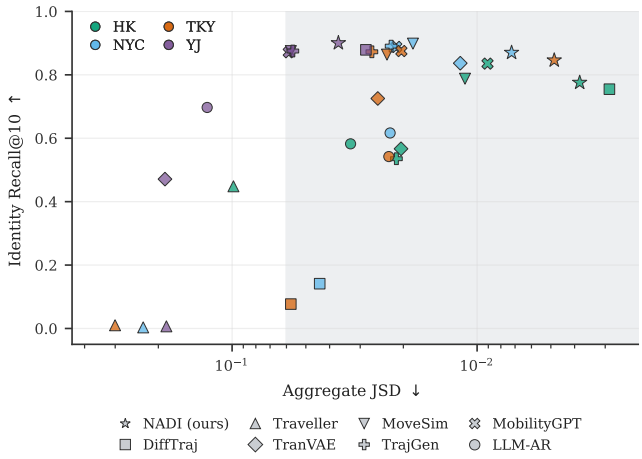

Figure 2: **Aggregate structural fidelity can conceal condition-specific identity loss.** Identity Recall@10 versus mean Jensen–Shannon divergence over six canonical mobility statistics from Feng et al. (2020) (lower is better; axis inverted), none of which captures per-user location identity.

We define the matched- and swapped-prefix conditions as

$$\xi_u = \left(T_u^{(K)}, N, w_{1:N}\right), \qquad \xi_{u\to u'} = \left(T_{u'}^{(K)}, N, w_{1:N}\right), \quad u' \neq u. \tag{7}$$

For each comparison, the reference future, sequence length, day-of-week scaffold $w_{1:N}$, and sampled corrupted state are held fixed; only the supplied user prefix changes. The resulting probe therefore isolates prefix-specific identity access at corruption level r, rather than completed-rollout behavior.

Let $\ell_i^\star$ denote the reference location at slot i. For pathway $\mathcal{P}$, we operationalize prefix-specific identity access as the matched-prefix advantage in probability assigned to the reference location:

$$\Gamma_{\mathcal{P}}(r) = \mathbb{E}\left[p_{\mathcal{P}}\left(\ell_i^\star \mid \widetilde{s}_{\mathcal{P},r}, \xi_u\right) - p_{\mathcal{P}}\left(\ell_i^\star \mid \widetilde{s}_{\mathcal{P},r}, \xi_{u\to u'}\right)\right]. \tag{8}$$

The expectation is over target test instances, source users, valid slots, and corruption realizations. Because the two conditions share the same sampled corrupted state, $\Gamma_{\mathcal{P}}(r)$ is a paired contrast.

A positive $\Gamma_{\mathcal{P}}(r)$ means that the matched prefix assigns more probability to the exact reference location than the swapped prefix. Because the contrast is computed within each pathway, latent and native corrupted states need not be directly comparable. The profile of $\Gamma_{\mathcal{P}}(r)$ across corruption levels instead traces how strongly target-user-specific location identity remains accessible during denoising.

Native-space absorbing endpoint. The diagnostic framework motivates the MASK endpoint, instantiated by NADI (*Native-space Absorbing Diffusion for Identity-preserving Mobility Generation*). It realizes the two interventions above by removing the learned compression interface and replacing Gaussian perturbation with absorbing-state masking in the native categorical space. Training and sampling follow masked discrete diffusion with confidence-parallel decoding (Chang et al., 2022; Sahoo et al., 2024; Shi et al., 2024). Appendix A details the diagnostic rationale for these choices, together with the corruption kernel, denoising objective, architecture, and decoding procedure.

4 Experiments

Experimental roadmap. We first establish that structural fidelity can conceal exact identity failure (§4.1). An ordered intervention then separates representation from corruption, while compression-stage audits test pre-diffusion aliasing and insufficient representational capacity (§4.2–§4.3). Prefix swaps test weak conditioning use, and timestep-wise probes trace the loss or instability of prefix-specific identity access during denoising (§4.4–§4.5). Additional validation of the native-mask endpoint is reported in Appendices H–K. Full experimental details are provided in Appendix C.

Table 1: Seven-day fixed-run results. OBJ is the four-axis composite; selected component metrics cover personalization, temporal structure, location fidelity, and diversity. Abbreviations follow Table 6; the full breakdown is reported in Appendix D.

Model	OBJ↑	Recall@10↑	Micro↓	Macro↓	LocV↓	GVG↓
HK ($\|\mathcal{V}\|$=91, smart-card)						
TransVAE	.919	.567	.031	.113	.0012	.019
MoveSim	.873	.788	.035	.094	.0008	1.57
TrajGen	.839	.536	.048	.113	.0021	3.45
MobilityGPT	.885	**.835**	.028	.083	**.0005**	1.34
LLM-AR	.852	.582	.066	.175	.0008	.748
DiffTraj	.911	.755	.0045	.022	.0049	.142
Traveller	.789	.448	.059	.144	.221	.848
NADI (ours)	**.953**	.776	**.0016**	**.016**	**.0005**	**.018**
NYC ($|\mathcal{V}|\approx$32K, check-in)						
TransVAE	.864	.837	.059	.220	.0072	.495
MoveSim	.849	**.898**	.029	.175	.0051	2.94
TrajGen	.843	.891	.037	.198	.0049	3.28
MobilityGPT	.853	.888	.030	.185	**.0043**	2.13
LLM-AR	.752	.617	.081	.211	.0371	3.70
DiffTraj	.791	.141	.0075	.081	.112	**.139**
Traveller	.614	.004	.081	.337	.225	1.17
NADI (ours)	**.921**	.870	**.0026**	**.066**	.0065	.217
TKY ($|\mathcal{V}|\approx$49K, check-in)						
TransVAE	.832	.725	.073	.284	.0116	.353
MoveSim	.819	.864	.048	.211	**.0043**	11.6
TrajGen	.834	.873	.042	.219	.0045	5.88
MobilityGPT	.821	**.875**	.049	.204	.0047	14.3
LLM-AR	.663	.542	.167	.339	.0410	5.01
DiffTraj	.776	.077	.0075	.060	.210	**.181**
Traveller	.564	.010	.114	.311	.364	39.4
NADI (ours)	**.926**	.846	**.0016**	**.041**	.0045	.199
YJ ($|\mathcal{V}|\approx$40K grid, GPS)						
TransVAE	.525	.471	.447	.693	.0249	156
MoveSim	.732	.875	.309	.551	.0105	427
TrajGen	.714	.875	.327	.577	.0107	323
MobilityGPT	.736	.871	.305	.541	.0114	430
LLM-AR	.767	.697	.156	.360	.0151	507
DiffTraj	.873	.879	**.0045**	**.037**	.0101	1.50
Traveller	.482	.007	.356	.682	.096	176
NADI (ours)	**.924**	**.900**	.0064	.059	**.0093**	.456

4.1 Can structural fidelity conceal exact identity failure?

Table 1 and Figure 2 reveal a clear decoupling between structural generative fidelity and exact per-user location identity. Table 1 reports the full four-axis OBJ, while Figure 5 isolates this tension by plotting Identity Recall@10 against Aggregate JSD over the non-identity structural axes. The desirable regime is therefore high Recall@10 with low Aggregate JSD. Across all four datasets, structurally plausible generation does not necessarily imply preservation of user-specific location identity. Several baselines attain low aggregate discrepancy while their identity recall deteriorates substantially, whereas NADI consistently remains in the high-recall, low-discrepancy regime. The strongest baseline varies by dataset, but NADI achieves the highest full OBJ on all four datasets, with fixed-run margins of $+.034$, $+.057$, $+.092$, and $+.051$; the corresponding three-seed margins range from $+.037$ to $+.103$ (Appendix C.5 and K.1).

This structure–identity decoupling is especially visible across two distinct representation pathways. Traveller denoises learned latent representations and retains Recall@10 of 0.448 on the 91-station HK vocabulary, but falls to 0.004–0.010 on the three large-vocabulary datasets. DiffTraj instead applies Gaussian diffusion to native-token embeddings. It remains structurally competitive—including low aggregate discrepancy on several datasets—yet its Recall@10 drops to 0.141 on NYC and 0.077 on TKY. Thus, a model can reproduce plausible population-level temporal and spatial statistics while generating the wrong locations for a particular user. Figure 5 establishes this failure pattern; the following experiments examine where in the generative pipeline the identity loss arises.

Table 2: Controlled representation–corruption comparison. The three core pathways are listed in the order of Equation (1); VQ and DiffTraj are separate controls. Pathway rows report the three-seed mean $\pm$ standard deviation of OBJ, with mean Recall@10 in parentheses. In the bottom block, each entry reports δ_k^{OBJ}, its corresponding ρ_k^{OBJ} in parentheses, and the Welch t-statistic after the semicolon, for $k \in \{\mathrm{rep}, \mathrm{mask}\}$. All quantities except the Welch t-statistics are defined in Equations (3)–(5).

Pathway (representation, corruption)	HK	NYC	TKY	YJ
Sequential intervention				
CONT+AE (AE latent, Gaussian)	.704±.011 (.58)	.578±.067 (.02)	.625±.070 (.07)	.580±.025 (.03)
GAUSS (native-token embedding, Gaussian)	.916±.002 (.63)	.895±.004 (.80)	.840±.010 (.63)	.750±.009 (.82)
MASK (NADI; native token, absorbing mask)	**.956±.002 (.78)**	**.897±.012 (.83)**	**.924±.001 (.84)**	**.938±.014 (.94)**
Additional controls				
VQ (VQ latent, absorbing mask)	.769±.003 (.55)	.682±.006 (.04)	.709±.004 (.14)	.684±.000 (.06)
DiffTraj (coordinate representation, Gaussian)	.911±.007 (.74)	.785±.008 (.12)	.774±.005 (.07)	.869±.004 (.87)
Sequential OBJ decomposition: δ_k^{OBJ} (ρ_k^{OBJ}); Welch t				
$k = \mathrm{rep}$	.212 (84%); 33.9	.317 (99%); 8.2	.215 (72%); 5.2	.170 (48%); 10.9
$k = \mathrm{mask}$	.040 (16%); 20.8	.002 (1%); 0.2	.084 (28%); 14.5	.187 (52%); 19.2

4.2 How is improvement distributed along the intervention path?

Table 2 evaluates the ordered intervention from Section 3. Replacing the learned autoencoder pathway with the native representation under Gaussian corruption accounts for 48–99% of the full-path OBJ improvement. On the three large-vocabulary datasets, it further accounts for 71–92% of the personalization-axis improvement, 75–93% of the VisitDist reduction, and 73–92% of the Recall@10 gain (Appendix E). Thus, most broad identity improvement occurs when learned compression is removed.

The masking step contributes unevenly across datasets and concentrates its gains on rare locations. It accounts for 1%, 16%, 28%, and 52% of the full-path OBJ improvement on NYC, HK, TKY, and YJ, respectively (Table 2). The Welch t-statistics reinforce this pattern: all adjacent-step OBJ changes yield large statistics ($t = 5.2$–33.9) except GAUSS→MASK on NYC ($t = 0.2$). Masking nevertheless accounts for 55–73% of the rare-location Recall@10 gain on HK, TKY, and YJ (Appendix E). VQ improves modestly over CONT+AE in OBJ but remains at or below 0.14 Recall@10 on the three large-vocabulary datasets (Appendix L). The ordered decomposition therefore links broad identity gains to removing learned compression and rare-location gains to absorbing masking; it localizes where performance changes, not the underlying mechanism.

4.3 Where does learned compression lose identity?

The ordered intervention localizes performance changes; compression-stage audits test whether exact-location distinctions are already degraded before diffusion. Definitions and additional results are provided in Appendix E.2.

Continuous autoencoding. Without diffusion, exact reconstruction accuracy is 97.6–100% overall and 95.1–99.8% on user-rare visits, but drops to 56.0–76.3% in the global frequency tail on the three large-vocabulary datasets (Table 13). Thus, globally infrequent locations are disproportionately affected. Yet 97.6–99.9% overall reconstruction accuracy coexists with only 0.010–0.066 downstream Recall@10 and near-zero prefix-specific identity access during latent denoising (Figure 3). Clean reconstruction fidelity therefore does not predict identity preservation under Gaussian latent generation.

Vector quantization. VQ ambiguity is already visible at quantization (Table 14): code-only majority-decoding accuracy is only 6.3–18.1% on the three large-vocabulary datasets. Many exact locations are therefore aliased before diffusion. Sequence context may help choose among locations sharing a code, but the code itself no longer uniquely specifies the original location.

Capacity control. Increasing the continuous latent width to $d_z = d_{\mathrm{model}}$ does not close the native-space gap on TKY or YJ; Recall@10 remains at or below 0.10 (Figure 7). Because this endpoint

Table 3: Output-level prefix intervention. Entries report the matched-prefix minus swapped-prefix change in per-user Recall@10; larger values indicate stronger dependence on the supplied user history. Paired user-level bootstrap 95% confidence intervals have a maximum half-width of 0.013.

	HK	NYC	TKY	YJ
cont+AE	−.001	+.004	+.011	−.000
gauss	+.225	+.782	+.554	+.749
mask = NADI	**+.330**	**+.852**	**+.765**	**+.877**

removes dimensional reduction while retaining the autoencoder pathway, limited latent width alone cannot explain the large-vocabulary failure.

Together, the audits locate VQ identity loss at quantization, whereas the continuous pathway retains high clean reconstruction accuracy but loses prefix-specific identity access during Gaussian latent generation.

4.4 Does generated identity depend on the correct user history?

The compression audits locate identity loss relative to the representation, but do not establish whether completed generations depend on the supplied user history. We therefore perform an output-level prefix intervention complementary to Eq. (8). For each target instance, we run the full generation process under the matched and swapped prefixes, while holding the reference future length N and day-of-week scaffold $w_{1:N}$ fixed. Source users are selected randomly or by matching activity level, with the same deterministic mapping used for all pathways. Table 3 reports the matched-prefix minus swapped-prefix change in per-user Recall@10.

The cont+AE pathway changes negligibly under the swap, whereas both native pathways exhibit large Recall@10 drops. Random and activity-matched sources produce nearly identical effects, making simple activity mismatch an unlikely explanation. Native generations also move closer to the source user's visitation profile for 99–100% of test windows, while cont+AE remains approximately equidistant from the target and source histories (Appendix F). Under the matched pathways tested here, native-space generation therefore retains substantially stronger output-level dependence on the supplied user history.

4.5 How does corruption affect prefix-specific identity access?

The output-level intervention measures dependence after generation; here we trace whether that dependence remains accessible during denoising using $\Gamma_{\mathcal{P}}(r)$ from Eq. (8).

Figure 3 follows denoising from $r = 1$ to the lowest evaluated corruption level, $r = 0.1$. MASK maintains strong and comparatively stable prefix-specific identity access throughout. GAUSS is markedly less stable: access rapidly vanishes on HK, emerges only transiently before declining on TKY and YJ, and remains substantial only on NYC, consistent with its small masking increment in Table 2. Thus, identity access does not necessarily recover as the state becomes cleaner. The CONT+AE pathway remains near zero across all evaluated corruption levels.

At $r = 1$, MASK also assigns higher probability and better rank to the reference location than GAUSS on all four datasets, whereas CONT+AE assigns approximately 10^{-4} probability on the large-vocabulary datasets (Appendix G). Since the pathways begin from different corrupted states, these endpoint comparisons are supporting rather than primary evidence. The trajectories instead expose two distinct failure modes: compression can suppress conditional identity access before denoising, whereas Gaussian corruption can make otherwise available access transient or unstable during denoising. Absorbing masking preserves a usable identity channel throughout generation, rather than merely improving the final output.

5 Limitations

Our mechanism claims are established within personalized mobility and the evaluated prefix-conditioned architectures; whether the same failure regime extends to other exact-entity tasks remains

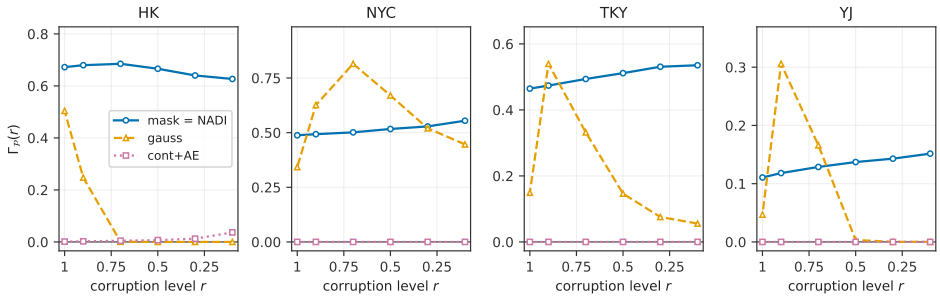

Figure 3: Prefix-specific identity access across evaluated corruption levels. The vertical axis is $\Gamma_{\mathcal{P}}(r)$, the paired matched-prefix minus swapped-prefix difference in probability assigned to the reference location; larger values indicate stronger access to user-specific location identity. The horizontal axis follows denoising from $r = 1$ on the left to $r = 0.1$ on the right. Panels use dataset-specific vertical scales.

open. The CONT+AE→GAUSS→MASK ladder is sequential rather than factorial, so it provides an order-dependent diagnostic decomposition rather than order-invariant causal effects. The VQ control also introduces representation-specific objectives and optimization choices. Hierarchical, compositional, or context-dependent codes, as well as alternative latent denoisers, may preserve identity differently. Accordingly, our results identify failure mechanisms in the tested pathways rather than establishing that learned compression must lose identity.

The diagnostic evidence also has finite scope. Latent-width and codebook-size interventions test insufficient capacity within the evaluated model families, but do not exhaust all possible capacity or optimization explanations. Several stage-specific probes use selected checkpoints, so their effect magnitudes lack multi-seed validation. Moreover, no dataset-level criterion yet predicts when native representation will provide the dominant gain or when absorbing masking will substantially improve over native Gaussian corruption.

The task supplies the reference future length N and slot-aligned day-of-week labels, and therefore evaluates reference-length-conditioned generation rather than autonomous visit counts, inactive days, or a fully calendar-consecutive future. NADI predicts over a dataset-specific closed vocabulary without explicit spatial geometry; DiffTraj's stronger held-out-user performance on YJ shows that native categorical prediction does not subsume coordinate-based spatial interpolation. Generalization to unseen locations, changing vocabularies, and new regions remains unresolved. We additionally evaluate vocabularies only up to approximately 49K locations and do not systematically study computational scaling, downstream simulation utility, memorization and privacy, sensing and discretization bias, or the unequal practical costs of different identity errors.

6 CONCLUSION

We studied when learned compression becomes a misaligned inductive bias for exact-entity conditional generation, where structurally plausible substitutes still constitute identity errors. An ordered representation–corruption intervention with stage-specific probes reveals two distinct mechanisms: vector quantization aliases exact locations before diffusion, while continuous autoencoding can preserve near-perfect clean reconstruction yet lose prefix-specific identity access during Gaussian latent denoising. Latent-width controls do not close the native-space gap, and prefix interventions expose the compressed pathway's weak dependence on the supplied user history. Removing learned compression accounts for most broad identity gains, whereas absorbing masking adds dataset-dependent improvements concentrated on rare locations and stabilizes identity access throughout denoising. NADI instantiates this diagnosis as native-space absorbing diffusion and validates it end to end across four heterogeneous datasets. More broadly, reconstruction fidelity and endpoint quality are not mechanism-identifying: for non-substitutable categorical outputs, representation and corruption should be judged by whether condition-supplied identity distinctions remain accessible throughout generation.

AI USE STATEMENT

Generative AI tools were used to assist with literature retrieval, research discussion and methodological feedback, drafting and language editing of parts of the manuscript, and refinement of presentation. All AI-assisted content, including technical statements and references, was reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work studies generative modeling of human mobility, which can involve privacy-sensitive behavioral data. We use the datasets solely for research under their respective access conditions; the HK smart-card data are accessed under a data-use agreement, while the remaining datasets are publicly available. Our notion of identity refers exclusively to exact task-defined location classes rather than personal identification. Nevertheless, models that preserve user-specific mobility patterns may raise privacy or memorization concerns, which are not systematically evaluated in this work and should be considered before deployment on sensitive mobility data.

REPRODUCIBILITY STATEMENT

All models follow a common prefix-conditioned task scaffold and are tuned independently by Bayesian optimization over family-specific search spaces, with configurations and checkpoints selected using validation rollouts. Appendices B and C document data construction, evaluation splits, model selection, metrics, and aggregation. The primary comparison between NADI and the strongest baseline on each dataset, all three rungs of the representation–corruption ladder, the selected VQ control, and the latent-width intervention are repeated over three training seeds. The remaining mechanism analyses use selected checkpoints and are reported as diagnostic.

The NYC and TKY Foursquare datasets[1] and the YJ dataset (YJMob100K[2]) are publicly available, whereas the HK smart-card data are available under a data-use agreement. Code, configurations, and evaluation scripts will be released upon publication.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://arxiv.org/abs/2107.03006.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025.

Xinyu Chen, Jiajie Xu, Rui Zhou, Wei Chen, Junhua Fang, and Chengfei Liu. Trajvae: A variational autoencoder model for trajectory generation. *Neurocomputing*, 428:332–339, 2021.

Chen Chu, Hengcai Zhang, and Feng Lu. Trajgdm: A new trajectory foundation model for simulating human mobility. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL)*, 2023. doi: 10.1145/3589132.3628362.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

[1] https://sites.google.com/site/yangdingqi/home/foursquare-dataset

[2] https://doi.org/10.5281/zenodo.10836269

Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. DeepMove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 World Wide Web Conference (WWW)*, pp. 1459–1468. ACM, 2018. doi: 10.1145/3178876.3186058.

Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. Learning to simulate human mobility. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 3426–3433. ACM, 2020. doi: 10.1145/3394486.3412862.

Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008. doi: 10.1038/nature06958.

Ada Görgün, Fawaz Sammani, Nikos Deligiannis, Bernt Schiele, and Jonas Fischer. Temporal concept dynamics in diffusion models via prompt-conditioned interventions. In *The Fourteenth International Conference on Learning Representations*, 2025.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Joshua Susskind, and Navdeep Jaitly. Matryoshka diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 30234–30262, 2024.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10696–10706, 2022. URL https://arxiv.org/abs/2111.14822.

Andi Han, Wei Huang, Yuan Cao, and Difan Zou. On the feature learning in diffusion models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7f7fa581cc8a1970a4332920cdf87395-Abstract-Conference.html.

Ammar Haydari, Dongjie Chen, Zhengfeng Lai, Michael Zhang, and Chen-Nee Chuah. MobilityGPT: Enhanced human mobility modeling with a GPT model, 2024. URL https://arxiv.org/abs/2402.03264.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pp. 13213–13232. PMLR, 2023.

Dou Huang, Xuan Song, Zipei Fan, Renhe Jiang, Ryosuke Shibasaki, Yu Zhang, Haizhong Wang, and Yugo Kato. A variational autoencoder based generative model of urban human mobility. In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 425–430. IEEE, 2019.

Wenjun Jiang, Wayne Xin Zhao, Jingyuan Wang, and Jiawei Jiang. Continuous trajectory generation based on two-stage GAN. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4374–4382, 2023. doi: 10.1609/aaai.v37i4.25557.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ANvmVS2Yr0. Oral presentation.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Yuxiao Luo, Songming Zhang, Kang Liu, Yang Xu, and Ling Yin. Traveller: Travel-pattern aware trajectory generation via autoregressive diffusion models. *Information Fusion*, 127:103766, 2026. doi: 10.1016/j.inffus.2025.103766. Available online in 2025.

Baichuan Mo, Zhan Zhao, Haris N Koutsopoulos, and Jinhua Zhao. Individual mobility prediction in mass transit systems using smart card data: An interpretable activity-based hidden markov approach. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):12014–12026, 2021.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pp. 7176–7185. PMLR, 2020.

Patrick Pynadath, Jiaxin Shi, and Ruqi Zhang. Candi: Hybrid discrete-continuous diffusion models. *arXiv preprint arXiv:2510.22510*, 2025.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q Tran, Jonah Samost, et al. Recommender systems with generative retrieval. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jinmeng Rao, Song Gao, Yuhao Kang, and Qunying Huang. LSTM-TrajGAN: A deep learning approach to trajectory privacy protection. In *11th International Conference on Geographic Information Science (GIScience 2021), Part I*, volume 177 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 12:1–12:17. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020. doi: 10.4230/LIPIcs.GIScience.2021.I.12.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022. URL https://arxiv.org/abs/2112.10752.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

Antonio Sclocchi, Alessandro Favero, Noam Itzhak Levi, and Matthieu Wyart. Probing the latent hierarchical structure of data via diffusion models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b503810343f1c7e80fab37b400dbc7a8-Abstract-Conference.html.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bad233b9849f019aead5e5cc60cef70f-Abstract-Conference.html.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. *arXiv preprint arXiv:2502.14831*, 2025.

Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010. doi: 10.1038/nphys1760.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://arxiv.org/abs/1711.00937.

Tonglong Wei, Youfang Lin, Shengnan Guo, Yan Lin, Yiheng Huang, Chenyang Xiang, Yuqing Bai, and Huaiyu Wan. Diff-RNTraj: A structure-aware diffusion model for road network-constrained trajectory generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):7940–7953, 2024. doi: 10.1109/TKDE.2024.3460051.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Qian Zhang, Lech Szymanski, Haibo Zhang, and Jeremiah D Deng. How reliable are semantic-id tokenizer comparisons in generative recommendation? *arXiv preprint arXiv:2605.25330*, 2026.

Tianxiong Zhong, Xingye Tian, Xuebo Wang, Xin Tao, and Pengfei Wan. Diffusing in the right space: A systematic study of latent diffusability. *arXiv preprint arXiv:2606.03578*, 2026.

Yuanshao Zhu, Yongchao Ye, Ying Wu, Xiangyu Zhao, and James Yu. Synmob: Creating high-fidelity synthetic gps trajectory dataset for urban mobility analysis. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, 2023a.

Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James J. Q. Yu. Difftraj: Generating gps trajectory with diffusion probabilistic model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023b.

Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Qidong Liu, Yongchao Ye, Wei Chen, Zijian Zhang, Xuetao Wei, and Yuxuan Liang. ControlTraj: Controllable trajectory generation with topology-constrained diffusion model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4676–4687, 2024. doi: 10.1145/3637528.3671866.

A NADI Formulation and Architecture

This section formalizes NADI (*Native-space Absorbing Diffusion for Identity-preserving Mobility Generation*), the native-representation and absorbing-corruption endpoint of the controlled diagnostic framework. For each instance, the conditioning scaffold is

$$\xi_u = \big(T_u^{(K)}, N, w_{1:N}\big),$$

where $T_u^{(K)}$ is the observed user prefix, N is the number of visits in the reference future, and $w_{1:N}$ are the corresponding slot-aligned day-of-week labels. Given this scaffold, NADI generates exactly N future visit tokens:

$$y_{1:N} = \big(y_i\big)_{i\in[N]}, \qquad y_i = (\ell_i, \tau_i, b_i) \in \mathcal{V} \times \mathcal{T} \times \{0, 1\}.$$

Here, ℓ_i, τ_i, and b_i denote the location, arrival-time, and day-boundary tokens at visit slot i, respectively. The sequence length N and day-of-week labels $w_{1:N}$ are supplied identically to all methods and are neither corrupted nor predicted. The model generates only the location, arrival-time, and day-boundary channels.

The generated boundary indicators partition the N visits into behavioral days:

$$d(i) = 1 + \sum_{j<i} b_j.$$

They determine the generated day structure, including the number of visits per day and the positions at which behavioral days end, but do not determine the sequence length. In particular, generation does not terminate when a boundary token is predicted; it terminates only after all N future slots have been completed.

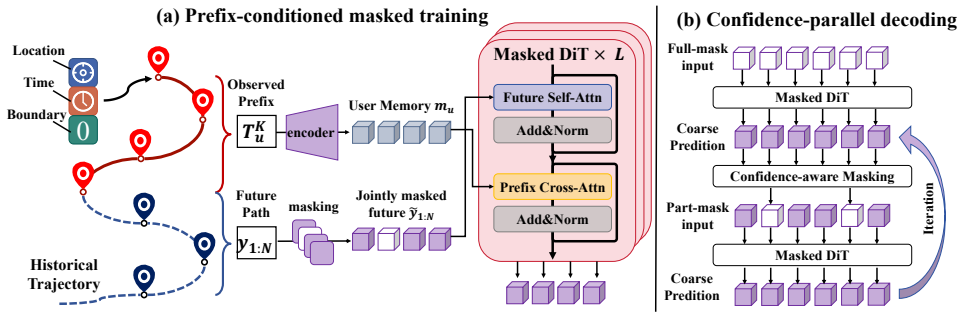

Figure 4: NADI architecture. (a) A masked DiT denoises jointly masked location, arrival-time, and day-boundary channels through cross-attention to the encoded prefix. (b) Starting from (N) fully masked future slots, confidence-parallel decoding commits high-confidence visits and jointly re-masks the rest. The observed prefix and slot-aligned weekday scaffold ($w_{1:N}$) remain fixed throughout decoding; the scaffold is omitted from the diagram for clarity.

Design rationale. NADI instantiates the native-mask endpoint of the diagnostic framework rather than asserting that discrete diffusion is generally superior to continuous diffusion. Its two design choices follow directly from the diagnosed failure modes. First, generation remains in the native categorical vocabulary, removing a learned representation interface through which exact location distinctions may be aliased or become difficult to recover. Second, absorbing corruption either retains an exact visit or replaces it with a mask state, rather than continuously perturbing its representation in an embedding or latent geometry. A masked identity must therefore be recovered directly from the supplied user prefix and sequence context. This keeps the native output ontology of exact location classes explicit throughout the generative pathway and makes condition-specific identity recovery the denoising task. These choices do not guarantee identity preservation, but remove learned compression and Gaussian corruption, the two design factors implicated by the diagnosis, thereby providing a controlled native-mask endpoint.

Native-vocabulary generation. NADI introduces neither an autoencoder nor a learned codebook. Locations are predicted directly over the dataset-specific vocabulary $\mathcal{V}$, while arrival time and day boundary are predicted through separate categorical heads over $\mathcal{T}$ and $\{0, 1\}$, respectively. Tokens are embedded internally for transformer computation, but each vocabulary item remains a distinct output class. Location identity is therefore never compressed into, quantized through, or reconstructed from a learned latent representation.

Absorbing-state corruption. Let

$$\boldsymbol{\mu} = \big([\mathrm{m}]_\ell, [\mathrm{m}]_\tau, [\mathrm{m}]_b\big)$$

denote the joint visit-mask tuple. For a corruption level $r \sim \mathcal{U}(0, 1)$, each future visit is independently retained with probability $1 - r$ or jointly replaced across all three generated channels with probability r:

$$q_r(\widetilde{y}_i \mid y_i) = (1 - r)\,\mathbf{1}\{\widetilde{y}_i = y_i\} + r\,\mathbf{1}\{\widetilde{y}_i = \boldsymbol{\mu}\}, \qquad i \in [N]. \tag{9}$$

Here, $\mathbf{1}\{\cdot\}$ denotes the indicator function. Under the standard monotone coupling across corruption levels, once a visit is masked, it remains masked as r increases. The process therefore either preserves an exact visit or hides it completely, rather than continuously perturbing its categorical identity. Joint masking preserves the alignment among the location, arrival-time, and day-boundary channels of each visit. The slot-aligned day-of-week labels $w_{1:N}$ are never corrupted.

Prefix-conditioned denoising. The observed prefix is encoded as sequence-valued user memory:

$$\mathbf{U}_u = E_{\mathrm{pre}}\big(T_u^{(K)}\big).$$

A noise-conditioned diffusion transformer denoises the corrupted future while cross-attending to this memory and using $w_{1:N}$ as deterministic calendar context:

$$(h_i)_{i\in[N]} = f_\theta\left(\widetilde{y}_{1:N}, r; \mathbf{U}_u, w_{1:N}\right). \tag{10}$$

The reference future length N is represented implicitly by the number of future slots supplied to the denoiser.

Three categorical heads predict the generated channels from the shared hidden state h_i:

$$\begin{aligned} &p_\theta\left(y_i \mid \widetilde{y}_{1:N}, r, \mathbf{U}_u, w_{1:N}\right) \\ &\quad = p_\theta(\ell_i \mid h_i)\, p_\theta(\tau_i \mid h_i)\, p_\theta(b_i \mid h_i). \end{aligned} \tag{11}$$

Although the output heads factorize conditionally on h_i, dependencies across channels and visit slots are represented through the shared transformer states, self-attention over the corrupted future, and cross-attention to the user prefix.

Training objective. Let

$$\mathcal{I} = \{i \in [N] : \widetilde{y}_i = \boldsymbol{\mu}\}$$

denote the masked visit positions. Corruption is implemented so that $\mathcal{I} \neq \varnothing$ for every training instance. The model minimizes the sum of the three channel-wise categorical cross-entropies, averaged over masked visits:

$$\mathcal{L}(\theta) = \mathbb{E}_{y,r,\mathcal{I}}\left[\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left(-\log p_\theta(\ell_i \mid h_i) - \log p_\theta(\tau_i \mid h_i) - \log p_\theta(b_i \mid h_i)\right)\right]. \tag{12}$$

In particular, the high-cardinality location head is trained directly with categorical cross-entropy over $\mathcal{V}$, without an autoencoder reconstruction objective or codebook commitment loss, following the simple objective used in masked discrete diffusion (Austin et al., 2021; Sahoo et al., 2024; Shi et al., 2024).

Confidence-parallel decoding. Generation begins with all N future visits replaced by the joint mask tuple:

$$\widetilde{y}_i^{(0)} = \boldsymbol{\mu}, \qquad i \in [N],$$

while the observed prefix and slot-aligned day-of-week scaffold remain fixed. At each MaskGIT-style decoding step, the model predicts all currently masked positions in parallel and commits the most confident visit predictions (Chang et al., 2022). Lower-confidence positions are jointly re-masked across location, arrival-time, and day-boundary channels and revised in subsequent steps. Decoding terminates when all N supplied future slots have been committed.

The resulting sequence is

$$\widehat{y}_{1:N} = \left(\widehat{\ell}_i, \widehat{\tau}_i, \widehat{b}_i\right)_{i\in[N]}.$$

The predicted boundary indicators $\widehat{b}_{1:N}$ then partition these fixed N visits into generated behavioral days. They determine the generated day organization but neither change the number of generated visits nor serve as a stopping condition.

B Visit-level task construction and relation to trip/OD pairs

B.1 Relation to adjacent mobility formulations

Personalized multi-day generation is designed to expose a failure mode that many adjacent mobility formulations do not isolate: a generated future may be structurally plausible yet contain the wrong exact locations for the conditioned user. Human mobility exhibits strong population-level regularities, while recurring locations remain highly user-specific (González et al., 2008; Song et al., 2010). We therefore treat location classes as exact, non-substitutable generated entities and condition future multi-day visit sequences on the user's recent history.

This formulation differs from single-step next-location prediction (Feng et al., 2018), which predicts one future state, and from mobility generators evaluated primarily for trajectory-level realism (Feng et al., 2020; Chu et al., 2023; Zhu et al., 2023b; Wei et al., 2024; Zhu et al., 2024; Luo et al., 2026), which do not necessarily isolate exact user-specific location identity. Our task instead evaluates categorical identity, temporal organization, distributional fidelity, and diversity jointly, while retaining a common conditional formulation for controlled representation–corruption comparisons.

B.2 BEHAVIORAL-DAY AND VISIT CONSTRUCTION

We convert all modalities into chronological visit sequences under a common behavioral-day convention. A behavioral day spans 04:00 on one civil date to 04:00 on the next and is labeled by the date on which it begins (Mo et al., 2021). Visits before 04:00 therefore inherit the preceding date's day-of-week label, and behavioral days without an observed visit are skipped.

For HK, each smart-card trip $(\text{in}, \text{out}, t_{\text{in}}, t_{\text{out}})$ is expanded into two visits, corresponding to tap-in and tap-out. Keeping these events separate avoids assuming that the exit station of one metro trip is the origin of the next: a user may leave the network at station Y and later re-enter at $Z \neq Y$. The resulting visit sequence therefore preserves the intervening off-network displacement and elapsed time. GPS grid observations and Foursquare check-ins are already point events and are used directly as visits. This conversion yields the common visit representation used in Section 2; dataset-specific discretization and sequence-length settings are reported in Appendix C.

B.3 WHY VISITS RATHER THAN TRIP/OD PAIRS

Origin–destination records are native to smart-card data, where both endpoints of each metro trip are observed. For GPS and check-in data, however, an OD pair must be constructed from two consecutive point observations. The same destination reached from different origins then becomes several distinct OD classes, although these transitions may express the same recurring location preference. An OD formulation therefore couples destination identity to the preceding origin and expands the event space from a location vocabulary to a pairwise origin–destination vocabulary.

Table 4: Per-user repetition of the most frequent origin–destination pair and location revisit rate. Location-level recurrence is substantially higher across all four datasets, supporting visits as the common generation unit.

dataset	modality	users	top-1 OD rep.	loc. revisit	days span
HK	smart-card (MTR)	29,237	32%	**94%**	∼120
YJ	mobile GPS (1 km grid)	97,408	5%	**89%**	75
NYC	Foursquare check-in	978	7%	**59%**	∼290
TKY	Foursquare check-in	2,091	7%	**65%**	∼290

C EXPERIMENTAL SETUP

This section specifies the datasets, common comparison scaffold, model-selection protocol, and evaluation metrics used throughout the experiments.

C.1 DATASETS AND PREPROCESSING

Table 5: Dataset statistics. *Coordinates* denotes the spatial representation; for HK, MDS denotes multidimensional scaling of network distances. $|\mathcal{V}|$ denotes the location-vocabulary size; *med. visits* is the median number of visits per user; and *test win.* is the number of eligible seven-day evaluation windows.

dataset	modality	coordinates	$\lvert\mathcal{V}\rvert$	users	med. visits	test win. (H=7d)
HK	smart-card tap (MTR)	network (MDS)	91	29,237	∼274	22,109
NYC	Foursquare check-in	lat/lon	∼32K	978	∼146	3,518
TKY	Foursquare check-in	lat/lon	∼49K	2,091	∼186	7,895
YJ	mobile-GPS 1 km grid	grid (x, y)	∼40K	97,408	∼880	23,899

Behavioral days start at 04:00 local time; a visit's day-boundary flag is 1 on the last visit of a day. Arrival times are binned to 96 fifteen-minute bins. Sequence length caps: 128 (HK/NYC/TKY), 352 (YJ).

C.2 Evaluation splits and windows

For the primary within-period evaluation, each instance pairs a seven-behavioral-day prefix with the user's next seven behavioral days. Training, validation, and test instances are constructed from the same dataset observation period; unlike the chronological evaluation, the primary split does not require all test windows to occur strictly after all training windows. Validation rollouts are used for hyperparameter and checkpoint selection, whereas the held-out test windows reported in Table 5 are used only for final evaluation. All methods receive the same prefix, reference future length N, and slot-aligned day-of-week context.

We separately evaluate stronger distribution shifts. The temporal-extension experiment (Appendix I) trains on the first 75% of each user's active days and evaluates on windows from the final 25%. The new-user experiment (Appendix J) applies the main-table checkpoints to users excluded from training. These settings test chronological extension and zero-shot user generalization, respectively.

C.3 Baselines and common generation protocol

The baseline suite includes continuous latent diffusion (Traveller (Luo et al., 2026) and our matched cont+AE control), coordinate diffusion (DiffTraj (Zhu et al., 2023b)), autoregressive and language-model generators (MobilityGPT (Haydari et al., 2024) and LLM-AR), adversarial models (MoveSim (Feng et al., 2020) and TrajGen (Jiang et al., 2023)), and TransVAE, a prefix-conditioned β-VAE adapted from trajectory and mobility VAEs (Chen et al., 2021; Huang et al., 2019; Higgins et al., 2017).

The mechanism study further includes native-Gaussian, native-mask, and vector-quantized controls (van den Oord et al., 2017; Gu et al., 2022). All methods receive the same prefix, target length, and day-of-week context, generate the same visit-level channels, and are selected under the same validation protocol. DiffTraj is adapted to the shared prefix-conditioned formulation in Appendix B.

C.4 Training, tuning, and checkpoint selection

Each model is tuned independently with Bayesian hyperparameter optimization over family-specific capacity, depth, and learning-rate choices, using 15 trials per model–dataset pair. Configurations and checkpoints are selected by the validation composite computed from sampled rollouts; test windows are never used for selection. All models use a seven-day prefix and seven-day horizon. The seeded main comparison, native-Gaussian control, and latent-width intervention use three training seeds. VQ, coordinate, identity-decomposition, prefix-swap, and posterior probes use a fixed checkpoint unless stated otherwise and are marked as diagnostic. All NADI models use $d_{\text{model}}{=}256$ and twelve parallel decoding steps.

dataset	d_{model}	depth	lr	val / test OBJ
HK	256	6	3.07×10^{-4}	0.9445 / 0.953
NYC	256	3	3.84×10^{-4}	0.9133 / 0.921
TKY	256	3	3.84×10^{-4}	0.9097 / 0.926
YJ	256	5	2.29×10^{-4}	0.8816 / 0.924

C.5 Evaluation metrics and the four-axis objective

The evaluation suite contains twelve metrics organized into four axes: fidelity, temporal structure, personalization, and diversity. The reported composite objective, denoted by OBJ $\in [0, 1]$ (higher is better), is computed from an eleven-metric objective set that excludes `DayOfWeek_DayShape_Cond_JSD` (DoWShape), which was available only as a post hoc diagnostic for the reported runs. For HK, `Spatial_KDE_JSD` (KDE) is additionally undefined. Metric availability is fixed at the dataset level and is identical across all models compared on that dataset.

For each dataset, OBJ is the equally weighted mean of the four available axis scores, with each axis score computed as the equally weighted mean of its available objective metrics. Fixed deterministic

transforms, rather than cross-model min–max normalization, map each raw metric to a $[0, 1]$ goodness score and make a single run's score comparable without reference to other configurations.

Base operator. Most metrics compare a real distribution p to a generated one q via the Jensen–Shannon divergence

$$\mathrm{JSD}(p\|q) = \tfrac{1}{2}\,\mathrm{KL}(p\|m) + \tfrac{1}{2}\,\mathrm{KL}(q\|m), \quad m = \tfrac{1}{2}(p+q), \tag{13}$$

computed with natural logarithms, so JSD $\in [0, \ln 2]$ (0 = identical distributions).

Metric provenance and notation. Abbreviations in all subsequent result tables follow Table 6. The fidelity metrics adapt distributional checks widely used in mobility and trajectory generation: LocV and ArrT compare visitation and temporal histograms (Feng et al., 2020; Rao et al., 2020), while KDE provides a representation-agnostic spatial-density counterpart to coordinate-based trajectory evaluation (Rao et al., 2020; Zhu et al., 2023a). Micro, Macro, and DoWShape are introduced here to measure within-day organization and day-of-week-conditioned routines, inspired by daily mobility motifs and periodic mobility regularity (Schneider et al., 2013; Feng et al., 2018). DoWShape is a complementary diagnostic of whether generated day shapes match the day-of-week scaffold provided identically to all models; it evaluates day-of-week-conditioned behavior rather than day-of-week prediction.

The personalization metrics adapt established analyses of individual visitation frequency, preferential return, and temporal regularity (González et al., 2008; Song et al., 2010; Feng et al., 2018). In particular, R@10 recasts top-K recall from mobility prediction as recovery of a user's frequent locations anywhere in the generated future. DShC, LocC, and GVG are introduced here as explicit support-coverage and per-user variance-collapse diagnostics, motivated by fidelity–diversity evaluation of generative models (Sajjadi et al., 2018; Naeem et al., 2020). Rare-loc R measures the recall of each user's rare ground-truth locations, defined as locations in the bottom half of the global frequency ranking, with no top-10 truncation. It is used only as an identity-sensitive diagnostic and is not included in OBJ.

Table 6: The full twelve-metric evaluation suite. DSHC and LOCC are reported as centered coverage gaps $\delta_{\mathrm{cov}} = \rho - 1$, where $\rho = C_{\mathrm{gen}}/C_{\mathrm{real}}$; OBJ applies the goodness transform to the uncentered ratio ρ. † Post hoc diagnostic; not included in OBJ. Spatial KDE is additionally unavailable for HK.

axis	metric (abbreviation)	reported quantity	dir.
A1 Fidelity	`LocVisit_TopK_JSD` (LocV)	JSD between top-K location-visit histograms, including an OTHER bin	↓
	`ArrTime_JSD` (ArrT)	JSD between arrival-time-of-day histograms	↓
	`Spatial_KDE_JSD` (KDE)	JSD between two-dimensional spatial kernel-density estimates	↓
A2 Temporal	`DayStruct_Micro_JSD` (Micro)	JSD over per-visit (within-day index, arrival-time bin) pairs	↓
	`DayStruct_Macro_JSD` (Macro)	JSD over per-day (#visits, first arrival, active span) tuples	↓
	`DayOfWeek_DayShape_Cond_JSD` (DoWShape)†	day-of-week-conditioned JSD between day-shape distributions	↓
A3 Personalization	`PerUser_VisitDist_JSD` (VisitD)	mean per-user JSD between real and generated location-visit distributions	↓
	`PerUser_VisitRecall@10` (R@10)	mean fraction of each user's ten most-visited real locations recovered in generation	↑
	`PerUser_TimeProfile_Cosine` (TimeP)	mean per-user cosine similarity between 24-bin time profiles	↑
A4 Diversity	`DayShape_Coverage_Ratio` (DShC)	centered day-shape coverage gap, $\delta_{\mathrm{cov}} = C_{\mathrm{gen}}/C_{\mathrm{real}} - 1$	→ 0
	`Loc_Coverage_Ratio` (LocC)	centered location coverage gap, $\delta_{\mathrm{cov}} = C_{\mathrm{gen}}/C_{\mathrm{real}} - 1$	→ 0
	`PerUser_GenVar_Gap` (GVG)	mean absolute gap in per-user variance of generated versus real visits per day	↓

Coverage-ratio reporting. For DShC and LocC, let

$$\rho = \frac{C_{\mathrm{gen}}}{C_{\mathrm{real}}}$$

denote generated coverage relative to the reference. Result tables report the centered gap

$$\delta_{\mathrm{cov}} = \rho - 1,$$

so $\delta_{\mathrm{cov}} = 0$ is ideal, with negative and positive values indicating under- and over-coverage, respectively. OBJ uses the uncentered ratio $\rho = \delta_{\mathrm{cov}} + 1$.

Per-metric goodness $g \in [0, 1]$. Each raw metric is mapped to a higher-is-better goodness score using a type-specific transform:

$$\text{JSD metrics:} \quad g = \max(0,\, 1 - v), \tag{14}$$

$$\text{recall/cosine } (v \in [0,1]): \quad g = \min(1, \max(0,\, v)), \tag{15}$$

$$\text{coverage ratio } (\rho > 0;\ \text{ideal } 1): \quad g = \max\!\left(0,\, 1 - \tfrac{1}{2}\left|\log_2 \rho\right|\right), \tag{16}$$

$$\text{variance gap } (v \geq 0): \quad g = \frac{1}{1+v}. \tag{17}$$

The coverage transform is symmetric on the log scale: ratios 0.5 and 2 both yield $g = 0.5$, while 0.25 and 4 yield $g = 0$. The variance-gap transform maps a perfect match, $v = 0$, to $g = 1$ and compresses unbounded positive gaps into $[0, 1]$.

Aggregation. For dataset d and axis a, let $\mathcal{M}^{\mathrm{obj}}_{a,d}$ denote the metrics included in the reported objective and available for that dataset. Let

$$\mathcal{A}^{\mathrm{obj}}_{d} = \left\{ a : \left| \mathcal{M}^{\mathrm{obj}}_{a,d} \right| > 0 \right\}$$

denote the corresponding set of nonempty axes. The axis score and overall objective are

$$s_{a,d} = \frac{1}{\left|\mathcal{M}^{\mathrm{obj}}_{a,d}\right|} \sum_{m \in \mathcal{M}^{\mathrm{obj}}_{a,d}} g_{m,d}, \qquad \mathrm{OBJ}_d = \frac{1}{\left|\mathcal{A}^{\mathrm{obj}}_{d}\right|} \sum_{a \in \mathcal{A}^{\mathrm{obj}}_{d}} s_{a,d}. \tag{18}$$

Metric availability is fixed at the dataset level and is identical across models evaluated on the same dataset. DoWShape was unavailable during objective computation for the reported runs and is therefore excluded from $\mathcal{M}^{\mathrm{obj}}_{a,d}$; values recomputed from cached rollouts are reported only as post hoc diagnostics. For HK, KDE is additionally undefined and is omitted from its fidelity-axis mean. The rank-stability analysis in Section K perturbs these equal axis and metric weights.

Welch t statistic. For any three-seed comparison, let $\{z^{(1)}_i\}_{i=1}^{n_1}$ and $\{z^{(2)}_j\}_{j=1}^{n_2}$ denote the seed-level composite scores of the two compared models or intervention rungs. We report

$$\Delta = \bar{z}^{(1)} - \bar{z}^{(2)}, \qquad t = \frac{\bar{z}^{(1)} - \bar{z}^{(2)}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

where $\bar{z}^{(k)}$ and s_k^2 are the sample mean and variance across seeds, with $n_1 = n_2 = 3$. The statistic allows unequal seed-level variances. A larger $|t|$ indicates that the observed mean difference is more stable relative to seed-level variation, rather than a larger effect size.

Table 7: Composite scores over three random seeds. The final row reports the mean difference and Welch statistic.

	HK	NYC	TKY	YJ
NADI (ours)	**.956±.002**	**.897±.012**	**.924±.001**	**.938±.014**
strongest baseline	TransVAE .912±.008	TransVAE .860±.018	TrajGen .821±.002	DiffTraj .869±.004
Δ (Welch t)	+.044 (9.0)	+.037 (3.0)	+.103 (81)	+.068 (8.0)

D Detailed component metrics for all models

Tables 8–11 report OBJ and the available objective metrics for the fixed single-seed runs underlying Table 1. OBJ follows the dataset-specific aggregation rule in Section C.5: DoWShape is excluded

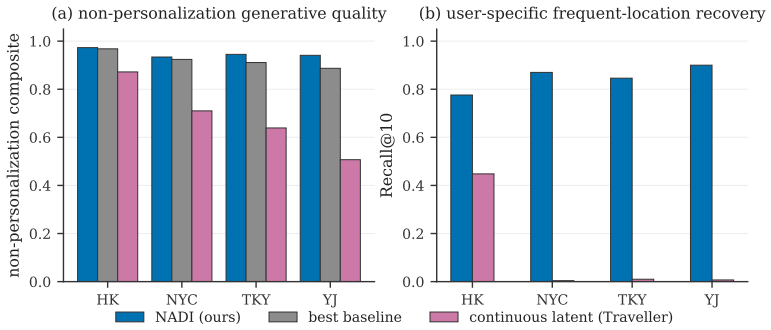

Figure 5: Non-personalization generative quality (left) and user-specific frequent-location recovery (right) for the fixed runs in Table 1. The left panel excludes the personalization axis from OBJ, while the right reports per-user Recall@10.

because it was computed only as a post hoc diagnostic, and spatial KDE is additionally unavailable for HK. Metric availability is fixed at the dataset level, so all models evaluated on the same dataset use the same objective metric set. Accordingly, DoWShape is omitted from the tables, while "—" denotes the unavailable KDE metric for HK. The DShC and LocC coverage ratios are reported after centering by one, so values closest to zero are best. Boldface denotes the best entry in each column; for DShC and LocC, this is the smallest absolute deviation from zero.

Table 7 separately reports the three-seed comparison between NADI and the highest-OBJ baseline on each dataset. The tables below provide the component-wise breakdown of the corresponding fixed single-seed runs and should not be interpreted as multi-seed estimates. The Figure 5 shows that NADI maintains non-personalization generative quality comparable to the best baselines while substantially improving user-specific frequent-location recovery across all four datasets.

Table 8: Detailed fixed-seed metric breakdown for HK ($|\mathcal{V}|$=91).

model	OBJ↑	LocV↓	ArrT↓	KDE↓	Micro↓	Macro↓	VisitD↓	R@10↑	TimeP↑	DShC→0	LocC→0	GVG↓
TransVAE	.9193	.0012	.0137	—	.0308	.1130	.0580	.5668	.8117	−.044	+.000	.0189
MoveSim	.8725	.0008	.0126	—	.0346	.0940	.0317	.7878	.8482	+.539	+.000	1.5669
TrajGen	.8391	.0021	.0194	—	.0475	.1126	.0619	.5355	.8115	+.262	+.000	3.4536
MobilityGPT	.8853	**.0005**	.0095	—	.0276	.0830	**.0254**	**.8354**	.8617	+.507	+.000	1.3351
LLM-AR	.8516	.0008	.0313	—	.0661	.1749	.0628	.5824	.6255	−.115	+.000	.7480
DiffTraj	.9106	.0049	.0033	—	.0045	.0217	.0984	.7547	.8733	+.805	+.000	.1424
Traveller	.7893	.2213	.0308	—	.0591	.1442	.5795	.4484	.7593	**−.020**	+.000	.8483
NADI	**.9530**	.0005	**.0008**	—	**.0016**	**.0160**	.0418	.7755	**.9474**	+.318	**+.000**	**.0181**

Table 9: Detailed fixed-seed metric breakdown for NYC ($|\mathcal{V}|$≈32K).

model	OBJ↑	LocV↓	ArrT↓	KDE↓	Micro↓	Macro↓	VisitD↓	R@10↑	TimeP↑	DShC→0	LocC→0	GVG↓
TransVAE	.8644	.0072	.0507	.0003	.0592	.2196	.1371	.8367	.6399	**−.047**	−.159	.4955
MoveSim	.8486	.0051	.0159	**.0002**	.0288	.1749	.0971	**.8984**	.7114	+.248	**−.125**	2.9383
TrajGen	.8428	.0049	.0193	.0002	.0369	.1981	**.0965**	.8905	.6936	+.200	−.128	3.2833
MobilityGPT	.8527	**.0043**	.0137	.0002	.0303	.1854	.1008	.8881	.7137	+.229	−.126	2.1328
LLM-AR	.7519	.0371	.0595	.0033	.0812	.2110	.3268	.6167	.6037	−.124	−.464	3.7025
DiffTraj	.7912	.1115	.0044	.0026	.0075	.0807	.6404	.1412	.6813	+.206	+.284	**.1392**
Traveller	.6136	.2246	.0443	.0721	.0805	.3374	.6903	.0035	.6588	−.132	−.928	1.1647
NADI	**.9210**	.0065	**.0003**	.0003	**.0026**	**.0657**	.1292	.8700	**.9038**	+.292	−.153	.2166

E Intervention and Latent-Pathway Audit Details

E.1 Identity-Level Intervention Decomposition

Table 12 evaluates the same sequential cont+AE→gauss→mask path on personalization-aware metrics. Results use a fixed checkpoint per rung and paired user-level bootstrap uncertainty; they complement the three-seed composite comparison in Table 2.

Table 10: Detailed fixed-seed metric breakdown for TKY ($|\mathcal{V}|\approx$49K).

model	OBJ↑	LocV↓	ArrT↓	KDE↓	Micro↓	Macro↓	VisitD↓	R@10↑	TimeP↑	DShC→0	LocC→0	GVG↓
TransVAE	.8317	.0116	.0638	.0007	.0726	.2835	.1776	.7254	.6154	−.208	−.180	.3532
MoveSim	.8188	**.0043**	.0239	.0003	.0475	.2110	.1171	.8638	.7126	+.270	−.158	11.6182
TrajGen	.8339	.0045	.0246	.0003	.0424	.2191	.1123	.8727	.7158	+.110	−.152	5.8749
MobilityGPT	.8205	.0047	.0252	.0002	.0489	.2036	**.1115**	**.8750**	.7125	+.268	**−.147**	14.3348
LLM-AR	.6630	.0410	.1228	.0037	.1670	.3392	.3943	.5422	.5577	−.434	−.551	5.0142
DiffTraj	.7757	.2102	.0050	.0051	.0075	.0596	.6697	.0769	.7064	+.179	+.342	**.1810**
Traveller	.5639	.3636	.0467	.1174	.1143	.3108	.6834	.0101	.6932	+.169	−.979	39.3634
NADI	**.9260**	.0045	**.0006**	**.0002**	**.0016**	**.0406**	.1432	.8460	**.9058**	+.204	−.160	.1987

Table 11: Detailed fixed-seed metric breakdown for YJ ($|\mathcal{V}| \approx 40$K).

model	OBJ↑	LocV↓	ArrT↓	KDE↓	Micro↓	Macro↓	VisitD↓	R@10↑	TimeP↑	DShC→0	LocC→0	GVG↓
TransVAE	.5250	.0249	.1962	.0008	.4474	.6930	.3219	.4711	.5358	−.952	−.470	156.1240
MoveSim	.7322	.0105	.0121	**.0003**	.3092	.5507	**.1825**	.8750	.9126	−.464	−.076	427.2595
TrajGen	.7145	.0107	.0138	.0003	.3273	.5770	.1826	.8752	.9043	−.554	−.083	322.8046
MobilityGPT	.7359	.0114	.0115	.0003	.3049	.5414	.1849	.8712	.9134	−.443	**−.076**	430.4248
LLM-AR	.7669	.0151	.0260	.0012	.1555	.3595	.2753	.6971	.8387	**−.006**	−.284	506.8669
DiffTraj	.8733	.0101	**.0014**	.0008	**.0045**	**.0366**	.3179	.8789	**.9402**	+.124	+.441	1.4977
Traveller	.4815	.0964	.0183	.0140	.3558	.6817	.6891	.0068	.9013	−.900	−.652	176.0963
NADI	**.9240**	**.0093**	.0033	.0004	.0064	.0587	.2051	**.9004**	.9224	−.008	−.131	**.4562**

Table 12: Identity-level decomposition along the ordered representation–corruption path. For each metric M and step $k \in \{\text{rep}, \text{mask}\}$, each entry reports δ_k^M, with its signed percentage of the full-path improvement ρ_k^M in parentheses, as defined in Equations (3)–(5). All metrics are expressed in higher-is-better form, so $s_M = +1$. Derived quantities are computed from unrounded values.

metric	step k	HK	NYC	TKY	YJ
R@10	rep	.107 (46%)	.781 (92%)	.567 (73%)	.772 (87%)
	mask	.126 (54%)	.072 (8%)	.213 (27%)	.118 (13%)
$1 - \text{VisitD}$	rep	.496 (95%)	.522 (93%)	.397 (75%)	.411 (85%)
	mask	.029 (5%)	.037 (7%)	.135 (25%)	.074 (15%)
A3	rep	.343 (85%)	.530 (92%)	.376 (71%)	.443 (84%)
	mask	.061 (15%)	.049 (8%)	.155 (29%)	.082 (16%)
Rare-loc R	rep	.098 (41%)	.525 (77%)	.298 (45%)	.062 (27%)
	mask	.143 (59%)	.159 (23%)	.368 (55%)	.164 (73%)

The rep and mask rows correspond to the two sequential steps in Equation (1). All metrics are averaged per user over the common eligible-user set. R@10 is the recall of each user's ten most-visited reference locations. VisitD is the mean per-user JSD between reference and generated visitation distributions; the table reports $1 - \text{VisitD}$. A3 is the equal-weight personalization score over R@10, $1 - \text{VisitD}$, and TimeP. Rare-loc R is the recall of reference locations in the bottom half of the global frequency ranking, without top-10 truncation.

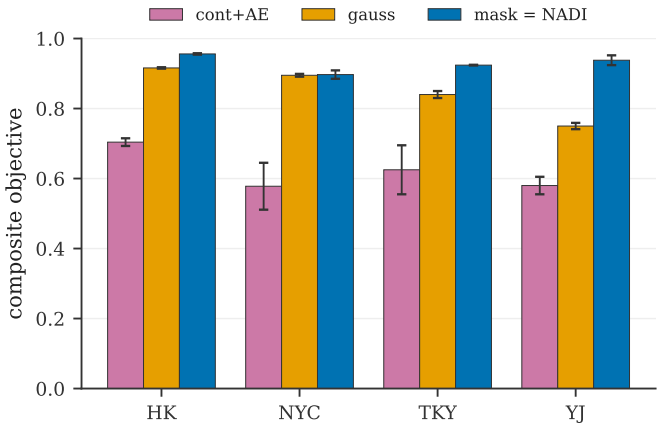

Figure 6: Sequential representation–corruption intervention. Replacing the learned autoencoder pathway with the native representation accounts for most of the composite recovery, while absorbing-state masking provides an additional dataset-dependent increment.

E.2 COMPRESSION-STAGE AUDIT DEFINITIONS

We audit the two compressed pathways before any generative corruption, using the selected checkpoint for each dataset. The *clean reconstruction audit* measures the exact-location reconstruction accuracy of the continuous autoencoder, both overall and within frequency-tail subsets (Table 13). The VQ code-ambiguity audit measures whether distinct locations are mapped to the same quantized code (Table 14). Together, the audits distinguish reconstruction error in the continuous pathway from location aliasing in the quantized pathway.

Continuous-autoencoder reconstruction. We disable diffusion and pass each observed visit through an uncorrupted encode–decode round trip. Let $\mathcal{D}_{\mathrm{audit}}$ denote the collection of observed audit visits, with each visit counted once. For $v \in \mathcal{D}_{\mathrm{audit}}$, let $\widehat{\ell}_{\mathrm{rec}}(v)$ denote its reconstructed location. For any audit subset $\mathcal{S} \subseteq \mathcal{D}_{\mathrm{audit}}$, the exact clean reconstruction accuracy is

$$R_{\mathrm{AE}}^{\mathrm{clean}}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \mathbf{1}\Big\{\widehat{\ell}_{\mathrm{rec}}(v) = \ell(v)\Big\}. \tag{19}$$

Let

$$n(\ell) = \sum_{v \in \mathcal{D}_{\mathrm{audit}}} \mathbf{1}\{\ell(v) = \ell\}$$

denote the global visit count of location ℓ, and write $w(v)$ for the evaluation window containing visit v. We evaluate Equation (19) on

$$\begin{aligned} \mathcal{S}_{\mathrm{all}} &= \mathcal{D}_{\mathrm{audit}}, \\ \mathcal{S}_{\mathrm{tail}} &= \{v \in \mathcal{D}_{\mathrm{audit}} : \ell(v) \in \mathcal{V}_{\mathrm{tail}}\}, \\ \mathcal{S}_{\text{u-rare}} &= \Big\{v \in \mathcal{D}_{\mathrm{audit}} : \textstyle\sum_{v' \in w(v)} \mathbf{1}\{\ell(v') = \ell(v)\} = 1\Big\}, \end{aligned} \tag{20}$$

where $\mathcal{V}_{\mathrm{tail}}$ is the dataset-level set comprising the bottom 10% of observed locations ranked by ascending global visit count $n(\ell)$, formed as equal-sized groups of locations rather than of visit mass. Thus, $\mathcal{S}_{\mathrm{tail}}$ tests globally infrequent location classes, whereas $\mathcal{S}_{\text{u-rare}}$ tests locations that the corresponding user visits exactly once within the evaluation window— one-off destinations rather than habitual anchors. The two strata are not nested and differ markedly in size: $\mathcal{S}_{\mathrm{tail}}$ covers 0.1%–2.4% of audit visits across datasets, whereas $\mathcal{S}_{\text{u-rare}}$ covers 4%–43%.

Table 13 reports $R_{\mathrm{AE}}^{\mathrm{clean}}(\mathcal{S}_{\mathrm{all}})$, $R_{\mathrm{AE}}^{\mathrm{clean}}(\mathcal{S}_{\mathrm{tail}})$, and $R_{\mathrm{AE}}^{\mathrm{clean}}(\mathcal{S}_{\text{u-rare}})$. Exact clean reconstruction accuracy is near-perfect overall and on user-rare visits, but drops sharply in the global frequency tail on the three large-vocabulary datasets. This pre-diffusion reconstruction error is nevertheless far smaller than the downstream collapse in Recall@10 and prefix-specific identity access. Clean reconstruction fidelity alone therefore does not explain the dominant identity failure during Gaussian latent generation.

VQ code-ambiguity audit. For a fixed dataset and VQ checkpoint, let $\widehat{p}$ denote the empirical distribution induced by all observed visits of the main users, with each visit counted once. For an observed visit $v = (\ell, \tau, b)$, define

$$L = \ell(v), \qquad \mathrm{AT} = \min\left\{\left\lfloor \frac{\tau(v)}{15} \right\rfloor, 95\right\} \in \mathcal{T} = \{0, \ldots, 95\}, \qquad Z = q\big(L, \mathrm{AT}, b(v)\big) \in \mathcal{Z},$$

where $\tau(v) \in \{0, \ldots, 1440\}$ is the arrival time in minutes, AT is the corresponding 15-minute arrival-time token, and $\mathcal{Z}$ is the VQ code vocabulary. The VQ encoder receives the location, arrival-time, and day-boundary channels, whereas the exogenous day-of-week label is excluded.

For each code $z \in \mathcal{Z}$, we define the empirical majority-location decoder

$$m(z) = \arg\max_{\ell \in \mathcal{V}} \widehat{p}(L = \ell \mid Z = z), \tag{21}$$

with ties broken deterministically. Its visit-weighted code purity is

$$\Pi_{\mathrm{VQ}} = \Pr_{\widehat{p}}[m(Z) = L] = \sum_{z \in \mathcal{Z}} \widehat{p}(Z = z) \max_{\ell \in \mathcal{V}} \widehat{p}(L = \ell \mid Z = z). \tag{22}$$

Thus, Π_{VQ} is the code-only majority-decoding accuracy. A low value indicates substantial location aliasing: many visits share codes whose majority location differs from their observed location.

As a secondary stress test, let U denote the user associated with a sampled visit. We define the set of highly user-specific locations as

$$\mathcal{V}_{\mathrm{spec}} = \left\{ \ell \in \mathcal{V} : \max_{u} \widehat{p}(U = u \mid L = \ell) \geq 0.8 \right\}. \tag{23}$$

Here, $\widehat{p}(U = u \mid L = \ell)$ is the share of visits to location ℓ contributed by user u. When $\mathcal{V}_{\mathrm{spec}} \neq \varnothing$, the conditional collision rate is

$$C_{\mathrm{VQ}}^{\mathrm{spec}} = \Pr_{\widehat{p}}[m(Z) \neq L \mid L \in \mathcal{V}_{\mathrm{spec}}]. \tag{24}$$

This quantity is undefined when no location satisfies the user-specificity threshold. To characterize which visit channels are preserved by the discrete code, for $A \in \{L, \mathrm{AT}\}$ we define the plug-in entropy and conditional entropy as

$$\begin{aligned} \widehat{H}(A) &= -\sum_{a \in \mathrm{supp}(A)} \widehat{p}(A = a) \log_2 \widehat{p}(A = a), \\ \widehat{H}(A \mid Z = z) &= -\sum_{a \in \mathrm{supp}(A)} \widehat{p}(A = a \mid Z = z) \log_2 \widehat{p}(A = a \mid Z = z), \\ \widehat{H}(A \mid Z) &= \sum_{z \in \mathcal{Z}} \widehat{p}(Z = z)\, \widehat{H}(A \mid Z = z). \end{aligned} \tag{25}$$

The corresponding mutual information is

$$\widehat{I}(A; Z) = \widehat{H}(A) - \widehat{H}(A \mid Z), \qquad A \in \{L, \mathrm{AT}\}. \tag{26}$$

All entropy and mutual-information quantities are maximum-likelihood plug-in estimates reported in bits. Accordingly, $\widehat{I}(L; Z)$ measures the average reduction in location uncertainty provided by the code, whereas $\widehat{I}(\mathrm{AT}; Z)$ measures the corresponding reduction in 15-minute arrival-time-token uncertainty. Because plug-in mutual information is upward biased in sparse contingency tables, these values should be interpreted as optimistic estimates of preserved information.

Table 14 shows severe code ambiguity on the three large-vocabulary datasets: code-only majority-decoding accuracy is only 0.063–0.181, while 97.5%–99.8% of visits to highly user-specific locations share a code whose majority label is different. Moreover, $\widehat{I}(\mathrm{AT}; Z) > \widehat{I}(L; Z)$ on every dataset, indicating that the codes preserve substantially more information about shared temporal regularities than about exact location identity. The low end-to-end R@10 values ($.05$–$.14$) show that this ambiguity persists into generation. HK is the partial exception, consistent with its much smaller location vocabulary.

Table 13: Clean reconstruction audit for the continuous autoencoder. The first three columns report exact clean location reconstruction accuracy over all audit visits, global-tail visits, and user-rare visits, respectively, using the subsets defined in Equation (20). The final column reports downstream per-user R@10 under Gaussian latent diffusion using the same checkpoint.

dataset	$R_{\mathrm{AE}}^{\mathrm{clean}}(\mathcal{S}_{\mathrm{all}})$	$R_{\mathrm{AE}}^{\mathrm{clean}}(\mathcal{S}_{\mathrm{tail}})$	$R_{\mathrm{AE}}^{\mathrm{clean}}(\mathcal{S}_{\text{u-rare}})$	R@10
HK	1.000	1.000	1.000	.542
NYC	.976	.660	.951	.017
TKY	.977	.560	.953	.066
YJ	.999	.763	.998	.010

E.3 Continuous latent-width intervention

We vary only the continuous latent width $d_z \in \{16, 32, 64, 128, 256\}$, keeping all other settings fixed to the selected *cont+AE* configuration and repeating each setting over three seeds. The intervention tests whether insufficient latent dimensionality alone explains the identity failure; it is not an independently re-tuned hyperparameter sweep. HK represents the small-vocabulary setting,

Table 14: VQ code-ambiguity audit. Mutual information is reported in bits; R@10 denotes end-to-end per-user Recall@10. "—" indicates that $\mathcal{V}_{\text{spec}} = \varnothing$.

dataset	$\lvert\mathcal{Z}\rvert$	Π_{VQ}	$C_{\text{VQ}}^{\text{spec}}$	$\widehat{I}(L;Z)$	$\widehat{I}(\text{AT};Z)$	R@10
HK	256	.394	—	2.59	4.36	.53
NYC	256	.063	.975	4.66	6.02	.05
TKY	2048	.181	.989	4.09	5.35	.14
YJ	4096	.063	.998	2.05	5.10	.04

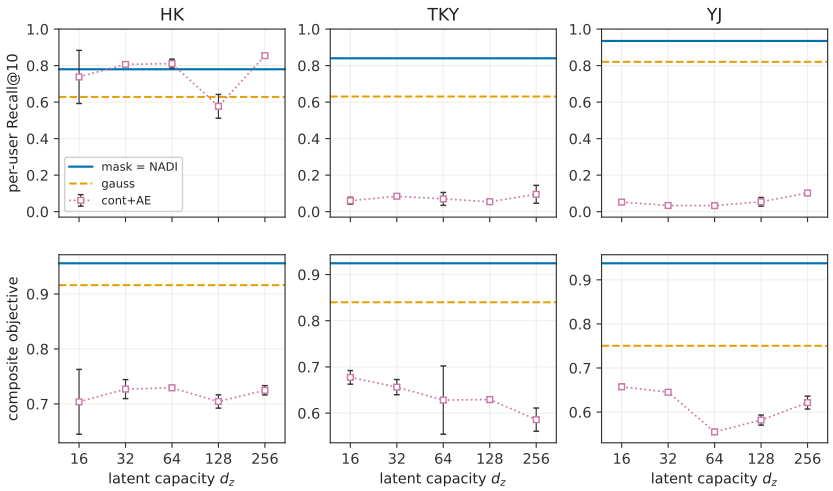

Figure 7: Continuous latent-width intervention on HK, TKY, and YJ. TKY serves as the representative large-vocabulary venue-check-in dataset; NYC is omitted because it has the same modality. Increasing d_z does not close the native-space identity gap on TKY or YJ. The rightmost point, d_z=d_{model}=256, removes dimensional reduction but retains the learned autoencoder pathway.

while TKY is used as the representative large-vocabulary check-in dataset; NYC is omitted because it shares the same modality. The endpoint $d_z = d_{\text{model}} = 256$ removes dimensional reduction but retains the learned autoencoder pathway.

Figure 7 and Table 15 show that increasing d_z does not close the native-space gap. On TKY and YJ, per-user Recall@10 remains at or below 0.10 across all widths, while the composite score shows no consistent improvement and remains well below both native pathways. HK is less severe: Recall@10 stays comparatively high and reaches 0.85 at $d_z = 256$, but the composite score remains around 0.70–0.73, substantially below *gauss* and *mask*. Thus, under the matched training protocol, limited latent width is not a sufficient explanation for the large-vocabulary identity failure.

Table 15: Continuous latent-width sweep: composite score (R@10), mean±std over three seeds.

pathway	HK	TKY	YJ
cont+AE, $d_z = 16$	.704±.059 (.74)	.677±.015 (.06)	.657±.006 (.05)
cont+AE, $d_z = 32$	.727±.017 (.81)	.656±.016 (.08)	.645±.004 (.03)
cont+AE, $d_z = 64$	.729±.004 (.81)	.628±.074 (.07)	.555±.005 (.03)
cont+AE, $d_z = 128$	.704±.012 (.58)	.629±.006 (.05)	.582±.011 (.05)
cont+AE, $d_z = 256$	.725±.008 (.85)	.586±.025 (.10)	.621±.015 (.10)
gauss	.916±.002 (.63)	.840±.010 (.63)	.750±.009 (.82)
mask = NADI	**.956±.002** (.78)	**.924±.001** (.84)	**.938±.014** (.94)

F PREFIX-SWAP ATTRIBUTION DETAILS

The same deterministic source-user mapping is used for all pathways. Source users are selected either randomly or by matching activity level; the two choices produce nearly identical effects. The

matched-prefix condition reproduces standard inference exactly, and Table 3 reports the corresponding R@10 intervention.

Window-level attribution. For test window m, let u_m denote the target user and $u'_m \neq u_m$ the source user whose prefix replaces the target prefix. The matched and swapped conditions are

$$\xi_{u_m} = \big(T^{(K)}_{u_m}, N_m, w_{1:N_m}\big), \qquad \xi_{u_m \to u'_m} = \big(T^{(K)}_{u'_m}, N_m, w_{1:N_m}\big).$$

The target future, sequence length, weekday scaffold, sampling seed, and evaluation window remain fixed; only the supplied prefix changes.

Let

$$\widehat{\pi}^{\mathrm{match}}_{\mathcal{P},m} = \widehat{\pi}_{\mathcal{P}}\big(\cdot \mid \xi_{u_m}\big), \qquad \widehat{\pi}^{\mathrm{swap}}_{\mathcal{P},m} = \widehat{\pi}_{\mathcal{P}}\big(\cdot \mid \xi_{u_m \to u'_m}\big)$$

denote the empirical location distributions of completed generations from pathway $\mathcal{P}$. Let π^{full}_u denote the empirical location distribution over user u's complete observed visitation record in the dataset. Thus, $T^{(K)}_u$ is only the seven-day conditioning slice and is not itself the distribution used in the JSD comparison.

For each window, we compute

$$J^{\mathrm{match|target}}_{\mathcal{P},m} = \mathrm{JSD}\big(\widehat{\pi}^{\mathrm{match}}_{\mathcal{P},m}, \pi^{\mathrm{full}}_{u_m}\big), \tag{27}$$

$$J^{\mathrm{swap|target}}_{\mathcal{P},m} = \mathrm{JSD}\Big(\widehat{\pi}^{\mathrm{swap}}_{\mathcal{P},m}, \pi^{\mathrm{full}}_{u_m}\Big), \tag{28}$$

$$J^{\mathrm{swap|source}}_{\mathcal{P},m} = \mathrm{JSD}\Big(\widehat{\pi}^{\mathrm{swap}}_{\mathcal{P},m}, \pi^{\mathrm{full}}_{u'_m}\Big). \tag{29}$$

Here, the first superscript term specifies the prefix condition used for generation, while the second specifies the user profile against which the generated distribution is compared.

Table 16 reports the window averages of these three quantities. It also reports the source-attribution rate

$$A_{\mathcal{P}} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}\Big[J^{\mathrm{swap|source}}_{\mathcal{P},m} < J^{\mathrm{swap|target}}_{\mathcal{P},m}\Big] \times 100, \tag{30}$$

where M is the number of test windows. Hence, $A_{\mathcal{P}}$ is the percentage of windows for which the swapped-prefix generation is closer to the source user's full visitation profile than to the target user's full profile. Ties are not counted as source-closer. Because this comparison is made within each window before aggregation, rounded mean JSDs may appear equal even when $A_{\mathcal{P}} \neq 50$.

Results. Under both native pathways, swapping the prefix makes the generated visitation distribution substantially closer to the source profile than to the target profile in 99.0–100% of test windows. Moreover, the swapped-to-source distance is close to the corresponding matched-to-target reference, indicating that the native pathways follow whichever user's prefix is supplied. In contrast, CONT+AE yields nearly identical distances to the target and source profiles and source-attribution rates of only 34–50%, indicating weak output-level dependence on the supplied prefix.

G GENERATION-START POSTERIOR DETAILS

At generation start, MASK receives an all-mask input, whereas GAUSS receives its highest-noise input. We report the mean probability assigned to the reference location under the matched-prefix condition. Because the pathways begin from different corrupted state spaces, this absolute endpoint comparison provides supporting evidence; the primary diagnostic of prefix-specific identity access is the within-pathway contrast $\Gamma_{\mathcal{P}}(r)$ across the corruption schedule in Figure 3.

H HORIZON GENERALIZATION

We evaluate the checkpoints trained at a seven-day horizon on rollouts of one, three, seven, and fourteen days. Among the models included at each horizon, NADI has the highest composite score in all sixteen dataset–horizon cells. DiffTraj is available only for the seven-day horizon in this sweep.

Table 16: Prefix-swap attribution. The three JSD columns report the window-averaged quantities in Eqs. (27)–(29); the final column reports the source-attribution rate $A_{\mathcal{P}}$ from Eq. (30). The reference profiles π_u^{full} and $\pi_{u'}^{\text{full}}$ are computed from each user's full observed visitation record; the 7-day prefix is used only as the conditioning input.

dataset	pathway	$\overline{J}_{\mathcal{P}}^{\text{match}\to\text{target}}$	$\overline{J}_{\mathcal{P}}^{\text{swap}\to\text{target}}$	$\overline{J}_{\mathcal{P}}^{\text{swap}\to\text{source}}$	$A_{\mathcal{P}}$ (%)
HK	CONT+AE	.59	.59	.59	50.1
	GAUSS	.11	.64	.11	99.9
	MASK = NADI	.09	.64	.10	100.0
NYC	CONT+AE	.69	.69	.69	33.8
	GAUSS	.34	.69	.37	99.6
	MASK = NADI	.33	.69	.37	99.6
TKY	CONT+AE	.68	.68	.68	42.7
	GAUSS	.39	.67	.40	99.0
	MASK = NADI	.33	.67	.37	99.4
YJ	CONT+AE	.69	.69	.69	41.5
	GAUSS	.34	.69	.34	99.9
	MASK = NADI	.28	.69	.29	100.0

Table 17: Generation-start posterior diagnostics for MASK = NADI / GAUSS (**bold** = better). The two pathways begin from different corrupted state spaces, so these endpoint quantities are reported as supporting diagnostics; the primary diagnostic of prefix-specific identity access is the within-pathway contrast $\Gamma_{\mathcal{P}}(r)$ in Figure 3. Larger KL divergence indicates greater departure from the population location marginal.

dataset	posterior on ref. location	KL→population	rank(ref. location)
HK	$\mathbf{0.692}/0.531$	$\mathbf{3.60}/2.81$	$1/1$
NYC	$\mathbf{0.480}/0.337$	$\mathbf{8.25}/7.72$	$\mathbf{1}/2$
TKY	$\mathbf{0.460}/0.147$	$\mathbf{7.15}/6.22$	$\mathbf{1}/4$
YJ	$\mathbf{0.113}/0.048$	$\mathbf{5.73}/4.96$	$\mathbf{6}/23$

Table 18: Composite objective vs. rollout horizon (same checkpoints; higher is better).

	HK				NYC			
model	H=1	3	7	14	H=1	3	7	14
TransVAE	.903	.912	.919	.904	.841	.826	.864	.878
MoveSim	.885	.872	.872	.896	.846	.864	.849	.880
TrajGen	.832	.857	.839	.857	.840	.862	.843	.872
MobilityGPT	.893	.888	.885	.906	.844	.868	.853	.889
LLM-AR	.826	.821	.852	.860	.770	.748	.752	.770
Traveller	.701	.755	.789	.787	.570	.588	.614	.647
NADI	**.950**	**.958**	**.953**	**.947**	**.890**	**.891**	**.921**	**.924**
	TKY				YJ			
TransVAE	.803	.801	.832	.847	.614	.522	.525	.534
MoveSim	.830	.838	.819	.856	.849	.772	.732	.740
TrajGen	.824	.843	.834	.875	.866	.763	.714	.724
MobilityGPT	.832	.841	.821	.846	.850	.774	.736	.744
LLM-AR	.679	.669	.663	.676	.762	.764	.767	.753
Traveller	.590	.573	.564	.567	.689	.545	.482	.458
NADI	**.865**	**.907**	**.926**	**.933**	**.888**	**.917**	**.924**	**.889**

Summary. NADI leads every reported dataset–horizon cell. On YJ, several one-shot baselines degrade markedly from one to seven days, whereas NADI remains stable. DiffTraj is available only at seven days and is therefore not part of the shorter- or longer-horizon comparison.

I TEMPORAL EXTENSION (GENUINE FUTURE)

For the chronological evaluation, models are trained on the first 75% of each user's active days and evaluated on windows from the final 25%. The prefix ends before the generated future window. This provides a stricter temporal-extension check than the main within-period evaluation. Windows: HK $33{,}587$ / NYC $4{,}699$ / TKY $10{,}968$ / YJ $36{,}123$ (all h=7d).

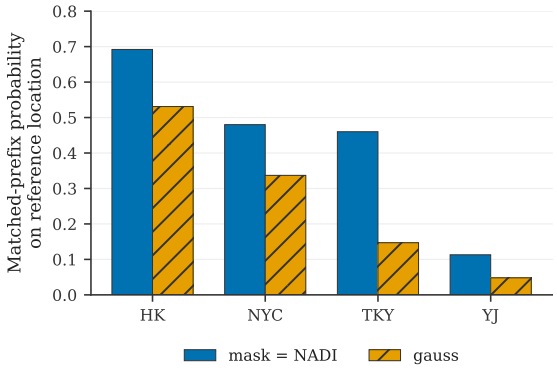

Figure 8: Mean model probability assigned to the reference location under the matched-prefix condition at generation start. MASK assigns higher reference-location probability than GAUSS on all four datasets. Because the pathways begin from different corrupted state spaces, this absolute endpoint comparison provides supporting evidence; the primary diagnostic of prefix-specific identity access is the within-pathway contrast $\Gamma_{\mathcal{P}}(r)$ across the corruption schedule.

Table 19: Temporal extension: composite objective on genuine-future windows.

model	HK	NYC	TKY	YJ
TransVAE	.912	**.786**	.776	.522
MoveSim	.834	.713	.713	.738
TrajGen	.823	.719	.729	.728
MobilityGPT	.835	.718	.710	.742
LLM-AR	.854	.699	.663	.748
Traveller	.780	.602	.547	.491
NADI	**.930**	.776	**.804**	**.914**

Summary. NADI leads on three datasets and changes only slightly on YJ (0.924 to 0.914). TransVAE narrowly leads on sparse NYC (0.786 versus 0.776), while Traveller remains lowest on all four datasets.

J NEW-USER (ZERO-SHOT) GENERALIZATION

Main-table checkpoints evaluated on $\sim$3,000 *held-out users* never seen in training (one deterministic 7d-prefix/7d-horizon window each); only HK and YJ have holdout users (26,230 / 93,663 beyond the main cohort). The A4 column is LocC-1 (signed deviation of location-coverage from ideal 1) because GVG is undefined with one window per user.

Summary. NADI leads on held-out HK users and among non-coordinate models on YJ. DiffTraj leads on held-out YJ (0.866 versus 0.816), consistent with the complementary spatial-interpolation bias of coordinates discussed in Section 5.

K ROBUSTNESS TO METRIC CHOICE AND STANDARD MOBILITY METRICS

We test whether the main ranking depends on the composite weighting and whether it persists under a separate suite of standard mobility statistics.

K.1 RANK-STABILITY UNDER RE-WEIGHTING

For each dataset d, let

$$\mathcal{M}_d^{\text{obj}} = \bigcup_a \mathcal{M}_{a,d}^{\text{obj}}$$

denote its reported objective metric set from Section C.5. Using the corresponding goodness transforms for all eight main-table models, we recompute NADI's rank under: (A) 20,000 random convex

Table 20: Zero-shot new users (representative columns; **bold** = best per column).

model	OBJ↑	R@10↑	Micro↓	Macro↓	LocV↓	LocC$-1\to 0$
HK new-user						
TransVAE	.8786	.5205	.0364	.1370	.0017	$+.000$
MoveSim	.8977	.5438	.0548	.1443	.0023	$+.000$
TrajGen	.8863	.5169	.0542	.1447	.0031	$+.000$
MobilityGPT	.8940	.5490	.0526	.1419	.0018	$+.000$
LLM-AR	.8447	.5098	.0695	.1885	.0014	$+.000$
DiffTraj	.8977	**.5897**	.0046	**.0323**	.0053	$+.000$
Traveller	.7674	.1592	.0584	.1540	.2226	$-.033$
NADI	**.9127**	.5620	**.0040**	.0337	**.0012**	$\mathbf{+.000}$
YJ new-user						
TransVAE	.4643	.1193	.4594	.6931	.0466	$-.584$
MoveSim	.6166	.2119	.2881	.5987	.0216	$-.272$
TrajGen	.5959	.2042	.3018	.6129	.0220	$-.289$
MobilityGPT	.6244	.2094	.2802	.5922	.0234	$-.284$
LLM-AR	.6987	.1680	.1666	.3968	.0414	$-.469$
DiffTraj	**.8664**	**.4930**	**.0045**	**.0993**	**.0126**	$+.329$
Traveller	.4495	.0009	.3656	.6846	.1066	$-.778$
NADI	.8161	.2702	.0104	.1279	.0273	$\mathbf{-.219}$

axis-weight vectors drawn from a Dirichlet distribution; (B) 20,000 random convex weight vectors over $\mathcal{M}_d^{\mathrm{obj}}$; (C) leave-one-axis-out; and (D) leave-one-objective-metric-out. Metric-level perturbations are restricted to the reported objective set: DOWSHAPE is excluded, and spatial KDE remains unavailable for HK.

Table 21: Rank stability of NADI under re-weighting of the reported objective set.

	HK	NYC	TKY	YJ
(A) random axis weights, $P(\mathrm{rank}=1)$	86.7%	99.5%	100%	98.4%
(B) random objective-metric weights, $P(\mathrm{rank}=1)$	88.5%	95.8%	99.3%	99.9%
(C) worst rank, leave-one-axis-out	1	1	1	1
(D) worst rank, leave-one-objective-metric-out	1	1	1	1
Equal-weight margin over second place (single seed)	+0.034	+0.057	+0.092	+0.051

NADI remains first in every leave-one-axis-out and leave-one-objective-metric-out evaluation. Random re-weighting changes the ranking mainly when nearly all weight is concentrated on a single diversity metric. The equal-weight configuration reproduces the fixed single-seed main table. This analysis measures sensitivity within the reported objective set; it does not test alternative metric definitions.

K.2 STANDARD MOBILITY-GENERATION METRICS

We additionally compare real and generated distributions over six standard mobility statistics: radius of gyration, trip or jump distance, global and per-user visitation rank, daily location count, and within-day duration (González et al., 2008; Song et al., 2010; Feng et al., 2020). Each discrepancy is measured by JSD, and none of these metrics is included in OBJ. Distances use venue coordinates for NYC and TKY, the 1-km grid for YJ, and an MDS embedding of the metro network for HK.

Table 22: Aggregate results across six standard mobility metrics and four datasets. Models are ranked within each metric–dataset cell. The final column counts cells with JSD above 0.1.

model	mean rank↓	mean JSD	median JSD	cells JSD>0.1
DiffTraj	3.29	0.0332	0.0047	3
MobilityGPT	3.79	0.0274	0.0155	1
MoveSim	3.83	0.0276	0.0158	1
TrajGen	4.17	0.0318	0.0205	0
LLM-AR	5.46	0.0512	0.0195	3
TransVAE	5.62	0.0614	0.0196	5
Traveller	7.17	0.2038	0.1680	15
NADI (ours)	**2.67**	**0.0132**	0.0067	**0**

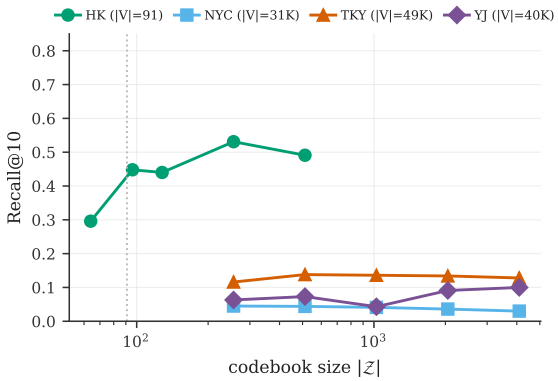

Figure 9: VQ codebook-size sweep. Increasing a single-code codebook does not recover personalized location identity on the large-vocabulary datasets.

L VQ-DIFFUSION CONTROL: DETAILS

The vector-quantized control in Table 2 is a full-visit VQ-Diffusion model (Gu et al., 2022) that maps each location, arrival-time, and day-boundary tuple to a single discrete code. It uses an EMA-updated codebook with dead-code reinitialization (van den Oord et al., 2017; Dhariwal et al., 2020) and is trained jointly with masked diffusion in code space, reconstruction loss, and commitment loss. This design matches the visit representation and downstream task of the continuous latent control, but necessarily introduces VQ-specific objectives and optimization choices. We search codebook sizes $\{64, 96, 128, 256, 512\}$ for HK and $\{256, 512, 1024, 2048, 4096\}$ for the three large-vocabulary datasets. The selected sizes are 256, 256, 2048, and 4096 for HK, NYC, TKY, and YJ, respectively.

Codebook-size sweep (Figure 9). With all other hyperparameters fixed, Recall@10 remains at or below 0.14 across all tested codebook sizes on NYC, TKY, and YJ. HK is a partial exception: recall improves once the codebook becomes comparable to its 91-location vocabulary and peaks at $|\mathcal{Z}| = 256$, but remains below the native-space models. Thus, enlarging the tested single-code-per-visit codebook does not recover user-specific location identity on the large-vocabulary datasets. This result does not rule out hierarchical, compositional, or context-dependent coding schemes.