The Scaling Bottleneck of Human Mobility Modeling

Lifeng Lin*
The University of Tokyo
Bunkyo, Tokyo, Japan
linlifeng@g.ecc.u-tokyo.ac.jp

Hangli Ge†
The University of Tokyo
Bunkyo, Tokyo, Japan
hanglige@g.ecc.u-tokyo.ac.jp

Xiang Zhang†
China University of Geoscience
Wuhan, Hubei, China
z.xiang@cug.edu.cn

Yao Yao
China University of Geoscience
Wuhan, Hubei, China
Hitotsubashi University
Kunitachi, Tokyo, Japan
yaoy@cug.edu.cn

Kazuma Hatano
The University of Tokyo
Bunkyo, Tokyo, Japan
kazuma.hatano@koshizuka-lab.org

Ryosuke Shibasaki
Reitaku University
Kashiwa, Chiba, Japan
LocationMind Inc.
Chiyoda, Tokyo, Japan
rshibasa@reitaku-u.ac.jp

Noboru Koshizuka
The University of Tokyo
Tokyo, Japan
koshizuka@iii.u-tokyo.ac.jp

Abstract

Although scaling—training larger models on larger datasets with more compute—has proven effective for gaining performance and generalization across many data modalities, a comparable trend has not yet emerged in human mobility modeling. Existing work in human mobility modeling has largely remained limited to relatively small models (<500M parameters) and modest data volumes (<1B samples), leaving it unclear whether, and how, scaling benefits this domain. This gap has limited progress toward human mobility foundation models for broad societal applications. In this work, we construct a large-scale real-world human mobility dataset comprising trillions of timestamped positioning records from millions of individuals, and take a first step toward scaling human mobility modeling. We first investigate whether human mobility models exhibit scaling behavior similar to that observed in other domains. Extensive experiments at data scales up to 10^{11} samples reveal an unexpected failure of conventional scaling in human mobility modeling. We then explore the source of this scaling failure and trace it to sample-quality issues inherent in the prevailing place-visitation (PV) modeling paradigm. Finally, we propose spatial-interaction behavior (SIB) modeling, a new paradigm designed to overcome this bottleneck and enable effective scaling in human mobility modeling. We show that SIB-based scaling recovers power-law scaling behavior and yields models that can be directly adapted to diverse downstream tasks. These findings identify both a key obstacle to scaling mobility models and a practical route toward scalable human mobility foundation models. Code and supplemental materials: https://fukamaru.github.io/human-mobility-scaling-bottleneck.

*Corresponding author. Work done during the internship at LocationMind Inc.
†Equal contribution.

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818976

CCS Concepts

• **Information systems** → **Geographic information systems**; • **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → **Machine learning**.

Keywords

human mobility, foundation models, neural scaling laws

ACM Reference Format:
Lifeng Lin, Hangli Ge, Xiang Zhang, Yao Yao, Kazuma Hatano, Ryosuke Shibasaki, and Noboru Koshizuka. 2026. The Scaling Bottleneck of Human Mobility Modeling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3770855.3818976

1 Introduction

Human mobility data can provide valuable insights into individual behavior[5], social phenomena [29, 30], and urban properties[41, 47, 50]. Modeling human mobility has therefore attracted multidisciplinary attention, ranging from urban management[36, 37] to epidemic prevention[1, 8, 9, 52]. Despite its promise, current human mobility models remain limited in performance and generalize poorly to unseen scenarios and diverse downstream tasks[31, 42]. The potential realization of mobility data remains at an early stage.

Meanwhile, progress in developing foundation models (*FMs*) has highlighted the effectiveness of scaling training[20, 21]. Benefits of scaling were first characterized in language modeling[20, 21] and have since been studied in vision[23, 43], time series[12, 13, 25, 55], and recommendation systems[18]. Across various modalities, larger models trained on more data samples consistently exhibit predictable performance improvement and emerging generalizability. Moreover, identified scaling laws provide quantitative guidance for model design, data allocation, and compute budgeting.

Human mobility modeling, however, has not yet been studied at comparable scales. Whereas FMs of other domains now involve

billions of parameters and massive training corpora, existing mobility models are typically much smaller and are trained on datasets with fewer than 10^9 GPS records. As a result, it remains unclear *how human mobility models behave at large scale, limiting progress toward scalable human mobility FMs.*

This work takes a first step toward answering this question. Scaling studies in human mobility have been difficult because metropolitan-scale longitudinal datasets are expensive to collect and subject to strict privacy constraints[51]. To address this challenge, we construct the *BlogWatcher (BW)* dataset, which contains time-stamped location records from millions of anonymized individuals over 48 months. This dataset provides the empirical basis for systematically studying how human mobility models behave as data and model scale increase, and for evaluating whether scaled models can generalize across settings and tasks.

Using the BW dataset, we first examine scaling behavior in human mobility modeling. We train model variances across a range of model and training set sizes. Contrary to the scaling trends observed in other domains, we find that increasing data scale does not produce the expected benefits. Instead, performance quickly plateaus as training data increase, suggesting that conventional scaling laws do not directly carry over to human mobility modeling. This scaling bottleneck suggests that increasing data and model scale alone is insufficient for developing human mobility FMs.

We attribute this bottleneck to the representational limitations of conventional place-visitation (PV) modeling. PV modeling represents individual mobility to a sequences of discrete location tokens. These design choices make sequence modeling poorly suited for scaling: *First*, location-token sequences are highly redundant due to the inherent nature of mobility behavior, allowing models to exploit frequency- or co-occurrence-based shortcuts rather than learning transferable behavioral structure[22]. *Second*, a location often conflates heterogeneous behaviors, activities, and populations, which can introduce conflicting learning signals during optimization. PV-based modeling thus tends to reach a performance plateau before additional data or parameters can be effectively converted into more generalizable representations.

To enable scaling human mobility modeling, we introduce Spatial-Interaction Behavior (SIB) modeling, a novel paradigm that represents mobility through informative spatial interactions rather than isolated location sequences. Instead of reducing trajectories to sparse and repetitive location tokens, SIB represents fine-grained GPS trajectories using two additional token types: Decision tokens and Search tokens. Decision tokens model stay behavior by representing time allocation at a location, whereas Search tokens model movement behavior through the direction and distance of movement toward subsequent locations. By representing both stays and transitions, SIB circumvent sample quality problem and retains behavioral and spatial structure that is often lost in location sequences, establishing a data foundation for scaling.

We instantiate this paradigm with *Riolu*, a Transformer-based architecture designed for efficient learning on SIB representations, to further address the efficiency problem in scaling mobility modeling. Our results show that SIB models recover predictable scaling behavior: performance improves as model size and training-data scale increase, in line with scaling trends observed in language and vision models. We further show that the pretrained SIB model transfers effectively to two downstream mobility tasks, suggesting that SIB provides a useful basis for general-purpose mobility representations. Finally, ablation studies identify modeling choices that are important for scaling SIB models, offering practical guidance for future large-scale mobility modeling.

The contributions of this work are summarized as follows:

- It is the first scaling study in human mobility domain.
- We identify a scaling bottleneck in conventional place-visitation modeling and analyze its connection to redundancy and spatial heterogeneity in mobility representations.
- We propose Spatial-Interaction Behavior (SIB) modeling and *Riolu* architecture to enable scalable mobility modeling in both data quality and computation efficiency.
- We show that SIB models recover predictable scaling behavior and achieve strong zero-shot transfer and state-of-the-art performance on downstream mobility tasks.

2 Background

2.1 Preliminaries

GPS trajectory. Mobility behavior of a user can be described via a GPS trajectory. For user u, we denote the trajectory by

$$\mathcal{T}_u = \big[(\ell_1, t_1), (\ell_2, t_2), \ldots, (\ell_n, t_n)\big],$$

where $\ell_i \in \mathbb{R}^2$ is the geographic coordinate and t_i is the corresponding timestamp. The trajectory is ordered chronologically, and the sampling intervals between consecutive records may vary.

Place-visitation representation. Most existing neural mobility models adopt a place-visitation (PV) representation, which converts $\mathcal{T}_u$ into a discrete token sequence. Specifically, a place mapping function $\phi : \mathbb{R}^2 \times \mathbb{R} \to \mathcal{V}_P$ assigns each observation or trajectory segment to a place token, yielding

$$\mathcal{P}_u = [v_1, v_2, \ldots, v_m], \qquad v_j \in \mathcal{V}_P.$$

Here, $\mathcal{V}_P$ is the place vocabulary, whose elements may correspond to grid cells, POIs, or administrative regions. Under this paradigm, human mobility is modeled analogously to language, with places treated as tokens. This formulation enables easy data processing and adapting standard sequence modeling objectives, but compresses informative mobility behavior into a sequence containing only visited place and their orders.

2.2 Related Works

Human Mobility Modeling. Human mobility modeling has been studied for decades, from early statistical models to recent neural approaches[31]. Deep learning has enabled models to learn mobility patterns from large-scale trajectory data and support applications such as transportation prediction, accident analysis, and location-based services[6, 35, 44]. Despite differences in architecture and supervision, most existing approaches represent mobility as sequences of place visits. This place-visitation paradigm is simple and widely used, but it abstracts away fine-grained movement structure and transition behavior. We argue that this abstraction becomes a major limitation when mobility models are scaled.

Foundation Models. Foundation models (FMs)[2] provide a general pretraining framework in which large models learn reusable representations from broad data sources and are then adapted to

downstream tasks. This framework has been successful in language and vision, and has motivated domain-specific FMs in areas such as biomedicine[19, 45, 49] and materials science[28]. These models suggest that large-scale pretraining can improve transfer performance and reduce dependence on task-specific labeled data.

Human mobility FMs, however, remain underdeveloped. Existing trajectory foundation models have focused mainly on vehicular mobility[57, 60], which does not fully capture the behavioral structure of human mobility. Human mobility involves not only movement between locations, but also stay behavior, dwell time, spatial context, and repeated interactions with places. Modeling these components jointly remains an open challenge and motivates the need for a mobility-specific foundation-model paradigm.

Neural Scaling Laws. Neural scaling laws describe how model performance changes with model size, data scale, and compute. First characterized in language modeling, such scaling behavior has since been examined across multiple domains, including vision[7, 23], time-series modeling, and biomedical applications. These studies show that scaling can provide quantitative guidance for model and data design. However, scaling is not determined by data volume alone. Empirical and theoretical studies suggest that data quality, redundancy, feature spectra[3, 4, 10, 27], and data-manifold geometry[38] all affect learning curves and scaling behavior. This perspective is particularly important for human mobility modeling: if the representation collapses trajectories into repetitive and semantically coarse location tokens, increasing data volume may not yield the expected scaling gains.

3 Scaling Human Mobility Dataset

Scaling human mobility models requires large-scale longitudinal data, which are difficult to obtain because of collection cost and privacy constraints. To support such a study, we construct the **BlogWatcher (BW)**, a large-scale human mobility dataset containing trillions of time-stamped GPS records from millions of users in the Greater Tokyo Metropolitan Area.

Data Collection. GPS data were collected by the data provider[1] through mobile applications from users who had granted location permissions and consented to data use. When an authorized application was active, positioning records were logged at certain intervals, associated with persistent de-identified user IDs. Before available, the data were processed using privacy-preserving procedures, including k-anonymization and differentially private aggregation where applicable. We did not receive any information that can identify specific individuals, and all analyses were conducted under the approved data-use protocol.

3.1 Preprocessing

We apply three preprocessing steps to improve data quality. *First*, we perform temporal continuity filtering by retaining users with at least 48 records per day and no inter-sample gap longer than 3 hours. *Second*, we apply geographic filtering to focus on the Greater Tokyo Metropolitan Area, selecting only users for whom more than 90% of records fall within the study region. *Third*, we use the TP algorithm[24] to detect and remove positioning errors. The final dataset covers approximately 1% of the population of the Tokyo Metropolitan Area with an average of 500M positioning records daily, providing a solid empirical basis for our scaling analysis. The details of processing and dataset statistics can be found in supplemental materials.

[1] https://www.blogwatcher.co.jp/

3.2 Training and Evaluation Datasets

Due to computational constraints, we use mobility data from 2018 to 2019 for training. This training corpus contains more than 200B GPS records, which is substantially larger than datasets commonly used in prior human mobility studies. We further construct an evaluation suite from the remaining *BW* dataset for comprehensive evaluations. Refer supplemental materials for details.

- **Standard benchmark.** We hold out a two-month window from the same user population used for training. This benchmark measures performance under temporal and population conditions similar to those seen during training.
- **COVID benchmark.** We evaluate on records from the same user pool during the COVID-19 pandemic, which help us understand model robustness to distribution shift, as pandemic restrictions substantially altered pre-pandemic mobility patterns.
- **Commuter benchmark.** We curate 1,000 unseen users with regular home–work commuting patterns. Data from these users provide a controlled evaluation to question whether the model learn the most common mobility pattern behavioral, rather than relying on user-specific trajectory memorization.

4 The Scaling Bottleneck of PV Modeling

Our primary goal is to elucidate the scaling behavior of human mobility modeling when data scale is no longer a barrier. Specifically, we investigate whether loss follows neural scaling laws and performance of pretraiend models, as feeding more data into the training of a large model. We focus on decoder-only Transformer. Several architectures such as DeepMove[14], despite their popularity in mobility modeling, are not considered because their hardware-inefficient backbones inherently hinder scalability in engineering. We train two model sizes, 100M and 500M parameters, across training-data scales from 10^9 to 10^{11} GPS records. Here, data scale denotes the number of GPS records used to construct the training sequences before tokenization. Further details on hyperparameters, optimizer, and training schedule are provided in supplemental materials.

We construct two scaling axes. In **population scaling**, we increase the number of users while keeping the observation window fixed to one year. In **temporal scaling**, we fix the cohort to 100,000 users and expand the period. This design separates the effect of adding more individuals from the effect of observing the same individuals for longer periods. Each experiment is run with five independent random seeds, and we report the mean result.

4.1 Empirical Findings

Anomalous scaling behavior. We observe that models trained on PV representation exhibits discrepant scaling behavior compared to established empirical findings, as shown in Figure 1. As training-data scale increases, test loss decreases only marginally, and the fitted scaling exponents for both the 100M and 500M models remain close to zero. Increasing model capacity does not change this trend. These results suggest that simply scaling training data and model

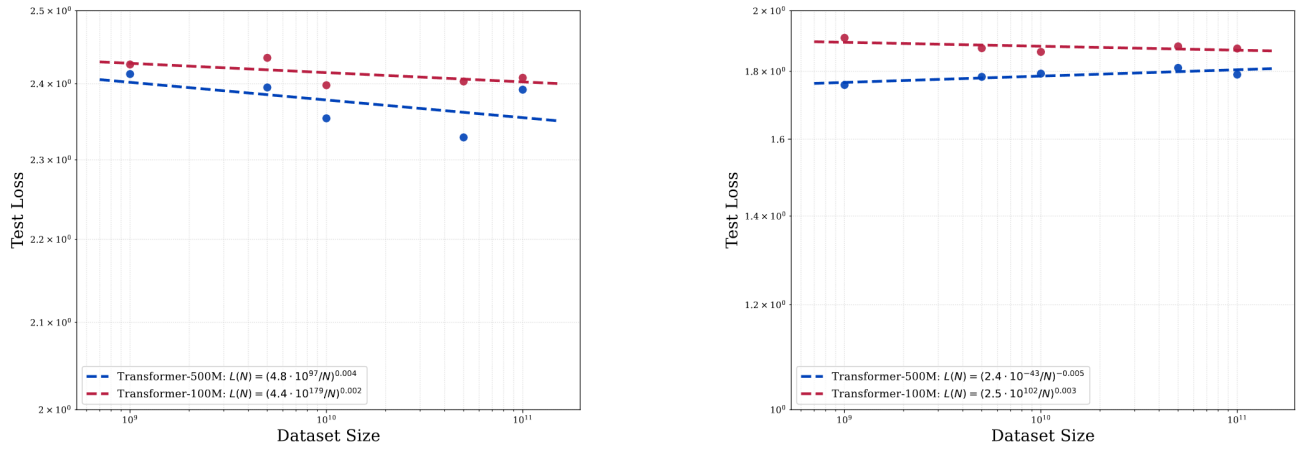


Figure 1: Scaling characteristics of training on the PV modeling paradigm. Test loss is shown as a function of dataset size N using a power-law fit $L(N) \propto N^{-\alpha}$. (*Left*) Scaling along the population dimension for a fixed time window. (*Right*) Scaling along the temporal dimension for a fixed population. The dashed lines represent the empirical fits for the 100M (red) and 500M (blue) parameter variants. In both dimensions, the scaling exponents α remain near zero ($\alpha \in [-0.005, 0.004]$), indicating that test loss remains largely invariant to increases in training data volume.

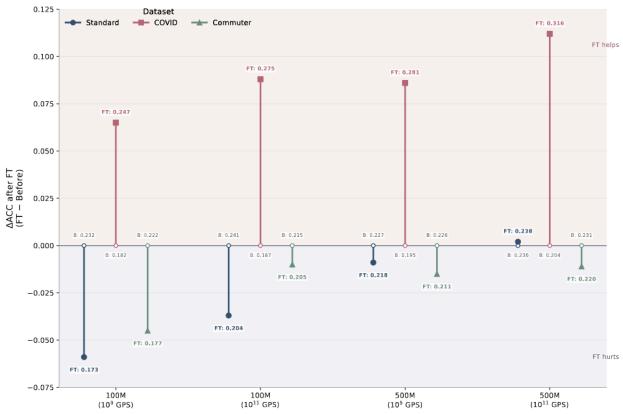


Figure 2: Fine-tuning effect on ACC across model scales and test sets. We report $\Delta\text{ACC} = \text{ACC}_{\text{FT}} - \text{ACC}_{\text{Before}}$, where positive values indicate improvement after fine-tuning. Hollow markers denote the before-FT baseline at zero, and filled markers denote the fine-tuned result. Text labels show the original before-FT and after-FT ACC values.

size fails to translate computation into test performance under conventional PV modeling paradigm.

Limited Downstream Improvement. The same pattern appears in next-location prediction. Although training data increase by orders of magnitude, prediction accuracy improves only slightly across evaluation settings. The limitation is most visible under distribution shift: on the COVID-period and the unseen commuters, larger models do not yield substantial gains. This indicates that scaling conventional PV modeling mainly improves fit to common mobility patterns, while offering limited benefits for shifted or specialized behavioral regimes.

Limited Knowledge Transferability. We also evaluate target-domain adaptation by fine-tuning the pretrained model on 10% of the COVID-period samples. The improvement on the COVID-period split is small, while performance on the standard benchmark decreases after fine-tuning. This trade-off suggests that the model adapts by fitting target-specific patterns rather than learning transferable mobility representations. Together, these findings motivate the need for a representation that preserves richer behavioral structure before scaling.

4.2 Why Scaling Fails: a Data-Centric Perspective

Our experiments suggest that the performance plateau observed during large-scale training cannot be explained by data scarcity alone. We argue that this lack of scalability is closely tied to the widely used place-centric paradigm, which represents mobility as sequences of visited locations. Although PV modeling is intuitive and computationally convenient, it introduces data quality issues.

Compression of trainable samples. The first issue is that conventional preprocessing substantially reduces the number and richness of trainable samples. Existing mobility modeling approaches typically construct samples through fixed-interval discretization, such as hourly time slots, or stay-point detection. Although high-frequency GPS trajectories may contain terabytes of spatiotemporal records, these preprocessing steps compress them into sparse place-visit sequences, discarding transient movement observations and within-interval variations.

Repetition-induced shortcut learning. In large-scale sequence modeling, excessive repetition can encourage memorization or shortcut learning, reducing generalization. Human mobility behavior is inherently repetitive[17, 33, 34]. As a result, converting GPS trajectories into place sequences produces samples dominated by a small set of frequently recurring locations. Such low-diversity

Table 1: Given the identical 200B GPS records, the application of distinct modeling paradigms yields sample sets that differ fundamentally in both quantity and quality.

Paradigm	# Samples	Compression Ratio
Place as Tokens	1.7B	0.153
SIB as Tokens	62.5B	0.604

sequences contain limited behavioral information beyond repeated place identities, allowing models to rely on frequency-based shortcuts rather than learn transferable mobility representations.

We provide further detailed discussion with empirical evidence in supplemental materials.

5 Toward Scalable Human Mobility Modeling

Crucial to addressing the scaling bottleneck, as highlighted by the previous evidence, is the reformulation of input representations that can encode rich behavioral and spatial information from GPS records while circumventing the sample-quality problem introduced by behavioral characteristics, thereby providing a solid data basis for scaling. Inspired by behavioral geography[16], we propose Spatial-Interaction Behavior (SIB) representation that implicitly preserve both common knowledge (e.g., urban structure and mobility patterns) and individual preference.

5.1 Spatial-Interaction Behavior as Tokens

In classical human mobility research, such as behavior geography and environmental psychology, human mobility is conceptualized as a spatial-interaction behavior, governed by a recursive decision-making process. Under this framework, mobility behavior is driven by shared social and subsistence needs, and in which locations function as spatial opportunities for need satisfaction. Individuals, therefore, continually interact with spatial opportunities by assessing how well each location satisfies their current needs, allocating time in proportion to perceived utility, and moving through space in search of alternative locations once the need is met or the current place becomes inadequate. We instantiate this theory by extracting SIB from raw GPS observations as tokens.

SIB representation. We define a Spatial-Interaction Behavior (SIB) representation is a chronologically-ordered sequence of SIB units. Each SIB unit b_j is a tuple

$$b_j = (L_j, D_j, S_j),$$

where L_j is a location token, D_j is a Decision token, and S_j is a Search token.

Decision token. Given a location token L_j, let $\mathcal{I} * j$ denote the set of GPS records associated with L_j. The Decision token is defined as

$$D_j = (t_j^{\text{arr}}, t_j^{\text{dep}}),$$

where

$$t_j^{\text{arr}} = \min *p_i \in \mathcal{I} * jt_i, \qquad t_j^{\text{dep}} = \max *p_i \in \mathcal{I}_j t_i.$$

Thus, $t_j^{\text{dep}} - t_j^{\text{arr}}$ measures the dwell duration associated with L_j. For a pass-through movement point, the arrival and departure timestamps are identical, yielding a zero-duration Decision token.

Search token. For two consecutive location tokens L_j and L_{j+1}, the Search token is defined as

$$S_j = (d_j, \Delta t_j, \theta_j) \in \mathbb{R}^3,$$

where d_j is the travel distance, Δt_j is the travel time, and θ_j is the azimuth from L_j to L_{j+1}. This token represents the observable transition between consecutive locations, implicitly encoding urban structures and cognitive map to navigate mobility behaviors.

Although latent constructs such as intent and spatial cognition are not directly observed in GPS records, SIB operationalizes them through observable behavioral proxies: dwell time reflects individual perception and utility evaluation of locations, while distance, travel time, and direction characterize the spatial relationship between locations.

5.2 From GPS to SIB Representation

Given a GPS trajectory, we first use the TP processing pipeline[24] to label each record as stay or move. Consecutive stay records are merged into one location token, and the first and last timestamps of the stay segment form its Decision token. Move records are retained as passing location tokens. For each passing location, the Decision token uses the same timestamp for arrival and departure, corresponding to zero dwell time. We then compute the Search token between each pair of consecutive location tokens using their distance, travel time, and azimuth. The resulting sequence alternates between location, Decision, and Search tokens. Compared with PV modeling, this extraction procedure preserves transition dynamics and yields more training units, as summarized in Table 1.

5.3 Riolu for SIB Modeling

We instantiate SIB with *Riolu*, a decoder-only Transformer architecture tailored to SIB sequences. *Riolu* differs from a standard Transformer language model in four components: multi-scale location tokenization, patch-level embedding, multi-patch prediction, and a continuous spatial loss.

Tokenization. An SIB unit consists of a location token, a Decision token, and a Search token. Decision and Search tokens are represented by continuous numerical features and are embedded with two separate linear projection layers. For location tokens, we avoid assigning a unique ID to every location. Instead, we introduce **multi-scale administrative tokenization**, in which each location is represented by administrative IDs at multiple spatial scales. Each scale has its own vocabulary and embedding layer, and the embeddings across scales are combined to form the final location embedding. This hierarchical design reduces vocabulary size and allows locations that share administrative regions to share higher-level spatial representations.

Patch embedding. Given a patch size N, *Riolu* groups every N consecutive SIB units into one patch. The features within each patch are concatenated and projected through a shared multilayer perceptron (MLP) to obtain a patch embedding. Patching shortens the effective sequence length, reducing the cost of self-attention, and provides local behavioral context before the sequence is processed.

Transformer backbone. The patch embeddings are processed by a decoder-only Transformer equipped with rotary positional embeddings (RoPE)[39] and Gated Attention[32]. FlashAttention[11] is

used to reduce attention memory cost and improve autoregressive inference efficiency. These components enable efficient training and evaluation on long mobility sequences.

Multi-patch prediction. Rather than predicting only the next token, *Riolu* predicts multiple future patches. This objective reduces the model's dependence on noisy local token transitions and encourages it to capture longer-horizon mobility regularities. The prediction loss is computed over all predicted future patches.

Continuous loss. SIB outputs include both continuous behavioral variables and spatial location targets. To avoid treating all incorrect locations equally, we formulate prediction as a regression problem. Each location is mapped to a continuous coordinate, and the model is trained to minimize the distance between predicted and ground-truth coordinates, together with regression losses for the continuous Decision and Search variables. This objective preserves spatial proximity in the loss: predictions close to the target location are penalized less than distant predictions.

6 Experiments

6.1 Scaling Laws of SIB Modeling

To evaluate the scalability of SIB, we use the same scaling setup described above. The main difference is that GPS trajectories are converted into SIB sequences and modeled with *Riolu*, the architecture designed for this representation.

Emergence of power-law scaling. As shown in Figure 3, the loss curves under SIB are well described by a power-law relationship with training-data scale. Test loss decreases monotonically with training-data size over orders of magnitude, indicating that SIB benefits consistently from additional data.

Sustained gains from model capacity. Increasing model capacity also improves performance. Riolu-500M achieves lower test loss than Riolu-100M across all training-data scales, suggesting that SIB modeling can benefit from increased model size without saturating within our tested range.

Population scaling provides larger gains. Comparing the two scaling axes shows that population scaling yields larger marginal gains than temporal scaling. The steeper decay exponent along the population-scaling axis suggests that adding new users provides more useful behavioral diversity than extending the observation window for the same users. This finding suggests that information provided by individual data may reach a saturation point. By contrast, new users introduce denser spatial interaction behavior with broader geographical range, empowering models to learn richer generalizable knowledge about urban space.

To evaluate whether scaling improves downstream utility, we apply the pretrained Riolu-500M model to two standard mobility tasks: location prediction (LP) and mobility generation (MG).

6.2 Location Prediction

Setup. The location prediction (LP) task is typically formulated as predicting the next visited location given a user's historical observations. We compare against representative baselines, including DeepMove[14], Flashback[53], GETNext[54], STAN[26], LSTPM[40], and Pretrained Mobility Transformer[48]. All baselines are trained on the same amount of data under the PV modeling paradigm. For comprehensive evaluation, we report top-k accuracy (*ACC@k*) and mean absolute error (MAE). While *ACC@k* measures whether the ground-truth location appears among the top-k predictions, MAE measures the geographic distance between the predicted and ground-truth locations. To examine whether performance is inflated by a home-location shortcut, we report metrics separately for all-day and daytime settings.

Results. Table 2 shows that Riolu consistently outperforms the baselines across benchmarks and metrics. It achieves the highest accuracy in most settings and yields the lowest spatial error, indicating improvements in both categorical prediction and geographic proximity. The improvement is especially pronounced on the Commuter benchmark, which can be a convincing evidence that Riolu capture this common mobility pattern. Riolu also maintains strong performance on the COVID-period and daytime settings, suggesting that scaling SIB pretraining improves robustness to distribution shift and reduces reliance on trivial home-location shortcuts.

6.3 Mobility Generation

Setup. Unlike the LP task, mobility generation (MG) aims to generate complete trajectories that are consistent with empirical mobility patterns. We compare against five MG baselines: MoveSim[15], ActSTM[56], TrajGAIL[58], DiffTraj[59], and LLMob[46]. Following prior work[46], we use Jensen–Shannon divergence (JSD) to measure the discrepancy between generated and empirical trajectories over four mobility-pattern distributions: step distance (SD), step interval (SI), daily activity routine distribution (DARD), and spatiotemporal visit distribution (STVD).

Results. Table 3 shows that Riolu better matches empirical mobility distributions than existing MG baselines in most settings. It achieves particularly strong results on DARD and STVD, suggesting that it captures daily activity routines and spatiotemporal visit patterns more accurately. Although LLMob performs competitively on SD, Riolu shows better alignment on global pattern related metrics. On the COVID-period, Riolu achieves a lower STVD than LLMob, indicating stronger robustness under changed mobility behaviors.

6.4 Ablation Study

We conduct ablation studies to assess the contribution of each design component in adapting Transformers to SIB representations.

Multi-scale administrative tokenization. As shown in Table 4, replacing multi-scale administrative tokenization with a vanilla place tokenizer reduces training efficiency because of vocabulary expansion. It also prevents representation sharing across locations, reducing prediction accuracy and increasing spatial error.

Patch embedding. Feeding individual SIB units directly into the Transformer increases the effective sequence length and reduces training efficiency. It also make the model more sensitive to inherent randomness and of human behavior observations, therefore reducing location prediction performance.

Multi-patch prediction. Similarly, training the model to predict only the next SIB patch reduces performance. Without multi-patch prediction, the model requires more autoregressive prediction steps, slightly reducing training efficiency.

Unified continuous loss. Removing the unified continuous loss improves training speed, but requires balancing classification and regression losses. This leads to weaker downstream performance.

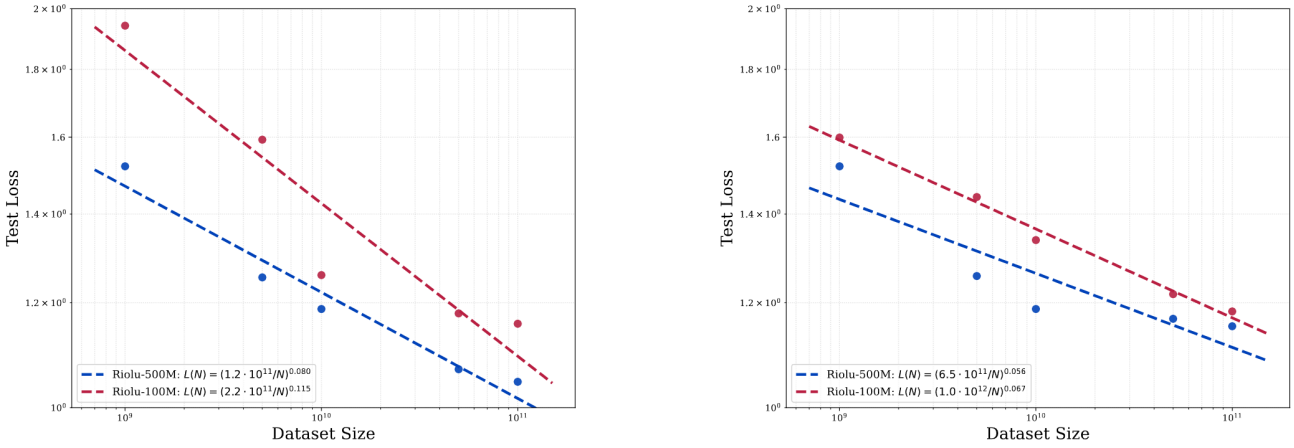


Figure 3: Scaling laws of SIB models across temporal and population dimensions. The plots illustrate the test loss as a function of dataset size (total GPS tokens) when scaling along the user axis (left, fixed observation duration) and the temporal axis (right, fixed user number). Dashed lines represent the power-law fits $L(N) = (C/N)^{\alpha}$. Both model variants exhibit robust scaling behavior, with temporal expansion yielding a higher rate of error reduction compared to population expansion

Table 2: Next location prediction (LP) performance results. A higher value indicates better performance for ACC@1 and ACC@5 while lower MAE in kilometers is preferred. The daytime, is defined as 7:30 AM to 7:30 PM.

Dataset	**Standard** (All Day / Daytime)			**COVID** (All Day / Daytime)			**Commuter** (All Day / Daytime)		
Metric / Method	ACC@1	ACC@5	MAE	ACC@1	ACC@5	MAE	ACC@1	ACC@5	MAE
DeepMove	0.149 / 0.102	0.443 / 0.396	5.81 / 12.03	0.170 / 0.111	0.343 / 0.225	4.17 / 16.42	0.236 / 0.177	0.490 / 0.563	7.22 / 15.26
Flashback	0.269 / 0.212	0.557 / 0.469	4.80 / **4.94**	0.338 / 0.127	0.394 / 0.267	**3.52** / 6.19	0.377 / 0.351	0.588 / 0.559	3.46 / 4.49
GETNext	0.204 / 0.145	0.473 / 0.305	8.05 / 13.35	0.254 / 0.118	0.396 / 0.275	6.19 / 16.82	0.132 / 0.126	0.551 / 0.488	5.71 / 11.98
STAN	0.171 / 0.152	0.402 / 0.349	9.69 / 15.83	0.224 / 0.174	0.330 / 0.231	7.62 / 14.37	0.236 / 0.192	0.285 / 0.236	7.47 / 10.98
LSTPM	0.198 / 0.159	0.413 / 0.366	8.57 / 9.41	0.227 / 0.146	0.452 / 0.306	6.45 / 13.79	0.253 / 0.174	0.397 / 0.253	9.08 / 12.13
PMT	0.272 / 0.131	0.558 / 0.441	7.85 / 8.63	**0.343** / 0.138	0.414 / 0.360	8.40 / 9.39	0.335 / 0.232	0.665 / 0.522	5.73 / 7.62
Riolu	**0.284** / **0.226**	**0.639** / **0.564**	**3.19** / 5.69	0.325 / **0.178**	**0.493** / **0.473**	3.69 / **4.79**	**0.434** / **0.368**	**0.737** / **0.696**	**3.27** / **4.01**

Table 3: Performance of mobility generation. Lower JSD is better. The best result is highlighted in bold, while the second-best is underlined.

Dataset	**Standard**				**COVID**				**Commuter**			
Metric / Method	SD	SI	DARD	STVD	SD	SI	DARD	STVD	SD	SI	DARD	STVD
MoveSim	0.043	0.181	0.296	0.507	0.115	0.063	0.178	0.494	0.069	0.125	0.134	0.489
ActSTM	0.090	0.108	0.273	0.403	0.217	0.084	0.402	0.609	0.063	**0.072**	0.134	0.481
TrajGAIL	0.117	**0.076**	0.238	0.553	**0.064**	0.091	0.289	0.711	0.055	0.120	0.373	0.315
DiffTraj	0.074	0.128	0.289	0.627	0.142	**0.046**	0.372	0.653	0.028	0.145	0.167	0.328
LLMob	**0.015**	0.136	0.193	**0.348**	0.096	0.072	**0.152**	0.547	0.059	0.082	0.110	**0.261**
Riolu	0.035	0.095	**0.144**	0.354	0.106	0.053	0.172	**0.338**	**0.024**	0.089	**0.098**	0.266

6.5 Roles of Decision and Search Tokens

We further analyze the roles of Decision and Search tokens through ablation studies on Riolu-500M. The results show that the two token types contribute complementary information: Decision tokens

Table 4: Ablation Study. All evaluation is conducted in next-location prediction task on the Standard Benchmark.

	Training Speed (s/iter)	ACC	MAE
w/o MSAT	2.183	0.165	5.59
w/o Patch Embedding	3.098	0.087	9.15
w/o Multi-patch Prediction	1.975	0.105	6.71
w/o Unified Continuous Loss	1.578	0.188	4.93
Standard Riolu	1.652	0.284	3.19

primarily encode user-specific spatial preference and cognition, whereas Search tokens help to capture the urban structure.

Decision tokens. A location identifier alone cannot distinguish the behavioral role of a place. The same location may serve as a workplace for one user, a restaurant for another, or a temporary stop for a third. Decision tokens provide an observable proxy for such differences by dwell time. Removing Decision tokens substantially degrades performance, with five of seven metrics dropping by more than 30% on the Commuter and COVID benchmarks. These two splits are particularly sensitive to spatial cognition and preference: commuters may share similar home–work routines but differ in the roles and durations associated with each location, while COVID-period mobility involves large shifts in dwelling preference, such as longer stays at home.

Search tokens. We believe Search tokens facilitate learning urban spatial information (e.g., distance and topology), which is lost during the discretization of continuous urban space but is implicitly embedded in human movement. Removing Search tokens mainly negatively affects spatial metrics, including MAE, SD, and STVD, indicating that transition information is important for both spatially accurate prediction and realistic mobility generation. These results confirm that SIB benefits from jointly modeling stay behavior and movement transitions.

6.6 Geographical Generalization

To test whether our findings extend beyond Tokyo, we conduct additional experiments in two regions with different urban characteristics: the Kansai metropolitan area (23B GPS records, 0.2M users, approximately 1.2% of the population) and Hokkaido (1.5B GPS records, 11K users, approximately 2.3% of the population). Kansai is another large metropolitan in Japan, whereas Hokkaido offers a lower-density and more geographically dispersed setting.

Scaling experiments. We repeat the scaling analysis in each region using fixed model sizes selected according to regional data scale and compute constraints: 300M parameters for Kansai and 100M parameters for Hokkaido. As training-data scale increases, PV modeling again shows little reduction in test loss, while SIB modeling exhibits clear scaling behavior. These results support our conclusion that the PV scaling bottleneck and the scalability of SIB are not confined to the Tokyo dataset.

Downstream tasks. We next evaluate whether the scaled SIB models improve downstream mobility tasks in the two additional regions. Figure 5 shows that SIB consistently improves downstream performance in both Kansai and Hokkaido, suggesting that its benefits persist across different regional settings.

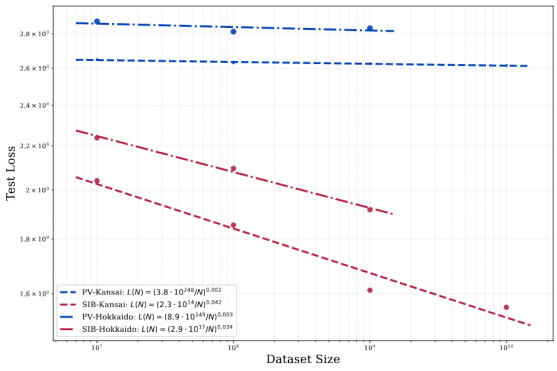
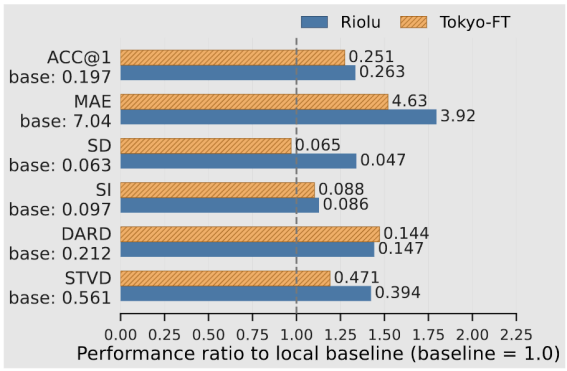
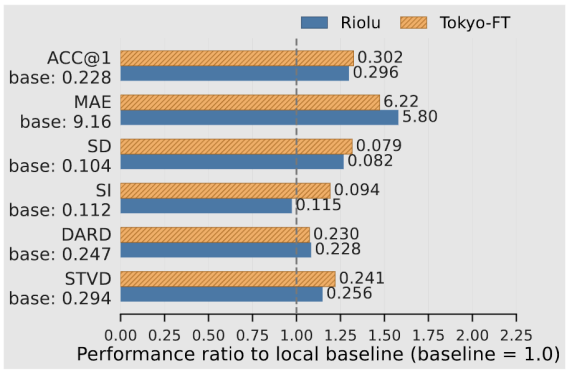


Figure 4: Scaling Experiments on Kansai and Hokkaido.



(a) The Kansai Metropolitan



(b) Hokkaido

Figure 5: Performance comparison on downstream tasks (LP and MG). Riolu refers to the Riolu model trained on SIB representations from Kansai/Hokkaido datasets, while Tokyo-FT is the fine-tuned Riolu pre-trained on BW Tokyo dataset.

Cross-region transferability. Finally, we test whether features learned from large-scale Tokyo pretraining can transfer to other regions. We adapt Riolu-Tokyo-500M to Kansai and Hokkaido by replacing the location embedding layers and coordinate prediction head, which are region-specific because of differences in location vocabulary and coordinate range. All remaining parameters are

Table 5: Ablation analysis of decision and search tokens. Each cells report the raw value with relative degradation against the full SIB representation in parentheses. For ACC, ↓ relative decrease is worse; for MAE, SD, SI, DARD, and STVD, ↑ relative increase is worse. Darker cells indicate degradation ≥ 50%, and lighter cells indicate degradation between 30% and 50%.

Metric	Standard			COVID			Commuter		
	Full	w/o decision	w/o search	Full	w/o decision	w/o search	Full	w/o decision	w/o search
ACC@1	0.284	0.268 (↓ 6%)	0.261 (↓ 8%)	0.325	0.279 (↓ 14%)	0.306 (↓ 6%)	0.434	0.355 (↓ 18%)	0.441 (↑ 2%)
ACC@5	0.639	0.578 (↓ 10%)	0.570 (↓ 11%)	0.493	**0.341 (↓ 31%)**	0.428 (↓ 13%)	0.737	**0.514 (↓ 30%)**	0.662 (↓ 10%)
MAE	3.19	**4.92 (↑ 54%)**	**5.78 (↑ 81%)**	3.69	**5.51 (↑ 49%)**	**5.16 (↑ 40%)**	3.27	**5.15 (↑ 57%)**	**5.35 (↑ 64%)**
SD	0.035	0.041 (↑ 17%)	**0.062 (↑ 77%)**	0.106	0.136 (↑ 28%)	**0.139 (↑ 31%)**	0.024	**0.038 (↑ 58%)**	**0.056 (↑ 133%)**
SI	0.095	0.115 (↑ 21%)	0.113 (↑ 19%)	0.053	**0.082 (↑ 55%)**	0.066 (↑ 25%)	0.089	0.115 (↑ 29%)	0.097 (↑ 9%)
DARD	0.144	**0.220 (↑ 53%)**	**0.195** (↑ 35%)	0.172	0.195 (↑ 13%)	0.184 (↑ 7%)	0.098	**0.129 (↑ 32%)**	0.110 (↑ 12%)
STVD	0.354	0.418 (↑ 18%)	**0.507** (↑ 43%)	0.338	**0.523 (↑ 55%)**	**0.588 (↑ 74%)**	0.266	0.316 (↑ 19%)	**0.393** (↑ 48%)

frozen. With only 10% of the target-region samples used for adaptation, Riolu-Tokyo-FT achieves competitive performance in Kansai and outperforms region-specific models in Hokkaido, where available training data are more limited. These results provide evidence that SIB-based pretraining learns mobility pattern knowledge with cross-region transferability.

7 Conclusion

We investigated whether human mobility modeling benefits from scaling training on metropolitan-scale longitudinal data. We find that conventional place-visitation modeling fails to benefit predictably from increased data and model scale. Our analysis suggests that this failure arises from the representation itself: reducing informative mobility behaviors to sequences of visited places creates sample-quality limitations that prevent additional data from yielding proportional gains. We address this bottleneck with Spatial-Interaction Behavior (SIB) modeling, which represents mobility through introducing decision tokens and search tokens beyond place tokens. We also propose Riolu architecture to circumvent computation efficiency problem to facilitate scaling training on SIB representations. Under this paradigm, we firstly recover predictable scaling behavior in human mobility modeling and achieve stronger transfer performance on downstream mobility tasks. Overall, our results show that scalable human mobility foundation models require not only larger datasets and novel architectures, but also representations that preserve the rich information of mobility behaviors.

8 Limitations and Ethic Consideration

Limitations. This work has several limitations. *First*, because BW is derived from smartphone applications, it may include demographic biases associated with smartphone ownership, app usage, and consent patterns. Older adults, minors, and other underrepresented groups may therefore be less well captured, which could affect model performance for these populations. *Second*, our analysis is restricted to three isolated region in Japan. Because large-scale mobility modeling requires substantial data and compute, future deployments may disproportionately benefit data-rich regions unless comparable resources are made available elsewhere. *Third,* compute and time constraints prevented us from fully exploring larger model sizes and pretraining on the complete dataset. *Finally*, current models does not incorporate external geographical and information, such as point-of-interest (POI) attributes and road networks. Incorporating such contextual signals may improve semantic representation and generalization across urban environments.

Ethic Consideration. The data used in this study were collected by BlogWatcher with users' consent for the collection and research use of de-identified location data. The data were processed before researcher access using privacy-preserving procedures, including k-anonymization and differentially private aggregation where applicable. None receive direct identifiers, such as names, phone numbers, or original device identifiers. All experiments were conducted in a restricted-access secure computing environment under the approved data-use protocol. Raw trajectory records remained within this environment, and only approved aggregate statistics, evaluation results, and model artifacts were exported. Given the sensitivity of fine-grained mobility data, we emphasize that mobility foundation models should be developed and deployed only with strict data governance, access control, privacy-preserving processing, and safeguards against misuse.

9 GenAI Disclosure

Generative AI tools were only used for language refinement to improve the presentation clarity. All intellectual context, research design and data analysis were conducted solely by the authors.

Acknowledgments

We thank Takashi Michikata for kind support and fruitful conversation on the early phase of this work. This study is partially supported by JSPS KAKENHI Grant Number JP25K21205.

References

[1] Alberto Aleta, David Martin-Corral, Ana Piontti, Marco Ajelli, Maria Litvinova, Matteo Chinazzi, Natalie Dean, M. Halloran, Ira Longini, Stefano Merler, Alex Pentland, Alessandro Vespignani, Esteban Moro, and Yamir Moreno. 2020. Modeling the impact of social distancing, testing, contact tracing and household quarantine on second-wave scenarios of the COVID-19 epidemic. doi:10.1101/2020.05.06.20092841

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On The Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258* (2021).

[3] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. 2024. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092* (2024).

[4] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. 2020. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*. PMLR, 1024–1034.

[5] Hancheng Cao, Fengli Xu, Jagan Sankaranarayanan, Yong Li, and Hanan Samet. 2020. Habit2vec: Trajectory Semantic Embedding for Living Pattern Recognition in Population . *IEEE Transactions on Mobile Computing* 19, 05 (May 2020), 1096–1108. doi:10.1109/TMC.2019.2902403

[6] Quanjun Chen, Xuan Song, Harutoshi Yamada, and Ryosuke Shibasaki. 2016. Learning deep representation from big and heterogeneous data for traffic accident inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2818–2829.

[8] Matteo Chinazzi, Jessica T Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, et al. 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* 368, 6489 (2020), 395–400.

[9] Shushman Choudhury, Abdul Rahman Kreidieh, Ivan Kuznetsov, and Neha Arora. 2024. Towards a trajectory-powered foundation model of mobility. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications*. 1–4.

[10] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM review* 51, 4 (2009), 661–703.

[11] Tri Dao. 2023. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. *ArXiv* abs/2307.08691 (2023). https://api.semanticscholar.org/CorpusID:259936734

[12] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*.

[13] Thomas DP Edwards, James Alvey, Justin Alsing, Nam H Nguyen, and Benjamin D Wandelt. 2024. Scaling-laws for large time-series models. *arXiv preprint arXiv:2405.13867* (2024).

[14] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. DeepMove: Predicting Human Mobility with Attentional Recurrent Networks. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1459–1468. doi:10.1145/3178876.3186058

[15] Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. 2020. Learning to Simulate Human Mobility. 3426–3433. doi:10.1145/3394486.3412862

[16] Reginald G. Golledge and Robert Stimson. 1996. Spatial Behavior: A Geographic Perspective. https://api.semanticscholar.org/CorpusID:145309517

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

[18] Xingzhuo Guo, Junwei Pan, Ximei Wang, Baixu Chen, Jie Jiang, and Mingsheng Long. 2023. On the embedding collapse when scaling up recommendation models. *arXiv preprint arXiv:2310.04400* (2023).

[19] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. 2024. Large-scale foundation model on single-cell transcriptomics. *Nature methods* 21, 8 (2024), 1481–1491.

[20] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556* (2022).

[21] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[22] Ryotaro Kawata, Yujin Song, Alberto Bietti, Naoki Nishikawa, Taiji Suzuki, Samuel Vaiter, and Denny Wu. 2025. From Shortcut to Induction Head: How Data Diversity Shapes Algorithm Selection in Transformers. *ArXiv* abs/2512.18634 (2025). https://api.semanticscholar.org/CorpusID:284078894

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4015–4026.

[24] Lifeng Lin, Hangli Ge, Takashi Michikata, Kazuma Hatano, Ryosuke Shibasaki, and Noboru Koshizuka. 2025. Robust and Efficient Human Mobility Data Processing through the Lens of Topological Persistence. *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems* (2025). https://api.semanticscholar.org/CorpusID:283747133

[25] Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. 2025. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816* (2025).

[26] Yingtao Luo, Qiang Liu, and Zhaocheng Liu. 2021. STAN: Spatio-Temporal Attention Network for Next Location Recommendation. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2177–2185. doi:10.1145/3442381.3449998

[27] Alexander Maloney, Daniel A Roberts, and James Sully. 2022. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859* (2022).

[28] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature* 624, 7990 (2023), 80–85.

[29] Esteban Moro, Dan Calacci, Xiaowen Dong, and Alex 'Sandy' Pentland. 2021. Mobility patterns are associated with experienced income segregation in large US cities. *Nature Communications* 12 (2021). https://api.semanticscholar.org/CorpusID:236531741

[30] Hamed Nilforoshan, Wenli Looi, Emma Pierson, Blanca Villanueva, Nic Fishman, Yiling Chen, John Sholar, Beth Redbird, David Grusky, and Jure Leskovec. 2023. Human mobility networks reveal increased segregation in large cities. *Nature* 624 (11 2023), 1–7. doi:10.1038/s41586-023-06757-3

[31] Luca Pappalardo, Ed Manley, Vedran Sekara, and Laura Alessandretti. 2023. Future directions in human mobility science. *Nature computational science* 3, 7 (2023), 588–600.

[32] Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free. *ArXiv* abs/2505.06708 (2025). https://api.semanticscholar.org/CorpusID:278502424

[33] Markus Schläpfer, Lei Dong, Kevin O'Keeffe, Paolo Santi, Michael Szell, Hadrien Salat, Samuel Anklesaria, Mohammad Vazifeh, Carlo Ratti, and Geoffrey B West. 2021. The universal visitation law of human mobility. *Nature* 593, 7860 (2021), 522–527.

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

[35] Xuan Song, Hiroshi Kanasugi, and Ryosuke Shibasaki. 2016. Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence*. 2618–2624.

[36] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, Teerayut Horanont, Satoshi Ueyama, and Ryosuke Shibasaki. 2013. Modeling and probabilistic reasoning of population evacuation during large-scale disaster. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1231–1239.

[37] Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, and Ryosuke Shibasaki. 2014. Prediction of human emergency behavior and their mobility following large-scale disaster. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 5–14.

[38] Stefano Spigler, Mario Geiger, and Matthieu Wyart. 2020. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment* 2020, 12 (2020), 124001.

[39] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2021. RoFormer: Enhanced Transformer with Rotary Position Embedding. *ArXiv* abs/2104.09864 (2021). https://api.semanticscholar.org/CorpusID:233307138

[40] Ke Sun, Tieyun Qian, Tong Chen, Yile Liang, Nguyen Hung, and Hongzhi Yin. 2020. Where to Go Next: Modeling Long- and Short-Term User Preferences for Point-of-Interest Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (04 2020), 214–221. doi:10.1609/aaai.v34i01.5353

[41] Xingye Tan, Bo Huang, Michael Batty, Weiyu Li, Qi Wang, Yulun Zhou, and Peng Gong. 2025. The spatiotemporal scaling laws of urban population dynamics. *Nature Communications* 16 (03 2025). doi:10.1038/s41467-025-58286-4

[42] Mark Tenzer, Zeeshan Rasheed, and Khurram Shafique. 2023. The Geospatial Generalization Problem: When Mobility Isn't Mobile. 1–4. doi:10.1145/3589132.3625590

[43] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems* 37 (2024), 84839–84865.

[44] Christopher Ververidis and G Polyzos. 2002. Mobile marketing using a location based service. In *Proceedings of the First International Conference on Mobile*

Business. Athens Greece, 1–12.

[45] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. 2024. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine* 30, 10 (2024), 2924–2935.

[46] Jiawei Wang, Renhe Jiang, Chuang Yang, Zengqing Wu, Makoto Onizuka, Ryosuke Shibasaki, and Chuan Xiao. 2024. Large Language Models as Urban Residents: An LLM Agent Framework for Personal Mobility Generation. *ArXiv* abs/2402.14744 (2024). https://api.semanticscholar.org/CorpusID:267782436

[47] Guangyuan Weng, Minsuk Kim, Yong-Yeol Ahn, and Esteban Moro Egido. 2025. Beyond Distance: Mobility Neural Embeddings Reveal Visible and Invisible Barriers in Urban Space. *ArXiv* abs/2506.24061 (2025). https://api.semanticscholar.org/CorpusID:280010811

[48] Xinhua Wu, Haoyu He, Yanchao Wang, and Qi Wang. 2024. Pretrained Mobility Transformer: A Foundation Model for Human Mobility. doi:10.48550/arXiv.2406.02578

[49] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature* 630, 8015 (2024), 181–188.

[50] Takahiro Yabe, Bernardo Garcia-Bulle, Morgan Frank, Alex Pentland, and Esteban Moro. 2023. Behavior-based dependency networks between places shape urban economic resilience. doi:10.21203/rs.3.rs-3683842/v1

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

[52] Chuang Yang, Zhiwen Zhang, Zipei Fan, Renhe Jiang, Quanjun Chen, Xuan Song, and Ryosuke Shibasaki. 2022. EpiMob: Interactive visual analytics of citywide human mobility restrictions for epidemic control. *IEEE Transactions on Visualization and Computer Graphics* 29, 8 (2022), 3586–3601.

[53] Dingqi Yang, Benjamin Fankhauser, Paolo Rosso, and Philippe Cudre-Mauroux. 2020. Location Prediction over Sparse User Mobility Traces Using RNNs: Flashback in Hidden States! 2156–2162. doi:10.24963/ijcai.2020/298

[54] Song Yang, Jiamou Liu, and Kaiqi Zhao. 2022. GETNext: Trajectory Flow Map Enhanced Transformer for Next POI Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1144–1153. doi:10.1145/3477495.3531983

[55] Qingren Yao, Chao-Han Huck Yang, Renhe Jiang, Yuxuan Liang, Ming Jin, and Shirui Pan. 2024. Towards neural scaling laws for time series foundation models. *arXiv preprint arXiv:2410.12360* (2024).

[56] Yuan Yuan, Jingtao Ding, Huandong Wang, Depeng Jin, and Yong Li. 2022. Activity Trajectory Generation via Modeling Spatiotemporal Dynamics. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 4752–4762. doi:10.1145/3534678.3542671

[57] Weijia Zhang, Jindong Han, Zhao Xu, Hang Ni, Hao Liu, and Hui Xiong. 2024. Urban foundation models: A survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6633–6643.

[58] Xin Zhang, Yanhua Li, Xun Zhou, Ziming Zhang, and Jun Luo. 2020. TrajGAIL: Trajectory Generative Adversarial Imitation Learning for Long-term Decision Analysis. *2020 IEEE International Conference on Data Mining (ICDM)* (2020), 801–810. https://api.semanticscholar.org/CorpusID:231626386

[59] Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James J. Q. Yu. 2023. DiffTraj: Generating GPS Trajectory with Diffusion Probabilistic Model. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:258298040

[60] Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Xun Zhou, Liang Han, Xuetao Wei, and Yuxuan Liang. 2024. Unitraj: Learning a universal trajectory foundation model from billion-scale worldwide traces. *arXiv preprint arXiv:2411.03859* (2024).