Marionette: Fine-Grained Conditional Generative Modeling of Spatiotemporal Human Trajectory Data Beyond Imitation

Bangchao Deng
University of Macau
Macao SAR, China
yc37980@um.edu.mo

Ling Ding
University of Macau
Macao SAR, China
mc36510@um.edu.mo

Lianhua Ji
University of Macau
Macao SAR, China
mc45079@um.edu.mo

Chunhua Chen
University of Macau
Macao SAR, China
chunhuachen@um.edu.mo

Xin Jing
University of Macau
Macao SAR, China
yc27431@um.edu.mo

Bingqing Qu
Beijing Normal-Hong Kong Baptist University, China
bingqingqu@uic.edu.cn

Dingqi Yang*
University of Macau
Macao SAR, China
dingqiyang@um.edu.mo

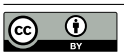
Abstract

Synthetic human trajectory data becoming increasingly prominent in various applications, including urban planning, traffic control, and crowd monitoring. Recent neural generative models for human trajectory data mostly follow an unconditional generative paradigm that relies on a pure data-driven imitative learning scheme, without considering the rich context of human mobility (e.g., social events or weather conditions) which may significantly impact the underlying human mobility patterns. Against this background, we propose Marionette, a **Ma**nipulatable gene**r**ati**i**ve m**o**del for human traje**c**t**o**ry da**t**a with fin**e**-grained conditions. Specifically, Marionette integrates both global and partial mobility-related contexts and extracts both sequence-level and event-level conditions. Afterward, it designs fine-grained and cascading conditioning mechanisms for modeling the temporal and spatial dynamics based on diffusion-alike Temporal Point Processes (TPPs) and discrete diffusion models, respectively, offering fine-grained controllable generative modeling of human trajectory data with both global and partial mobility-related contexts. We conduct a thorough evaluation on two real-world human trajectory datasets against a sizeable collection of baselines. Results show that our Marionette consistently outperforms the best baselines by 13.96-54.13% on statistical and distributional similarity metrics and by 9.36-40.63% in task-based data utility evaluation. Ablation studies verify our key design choices. Case studies also demonstrate the manipulability of Marionette in generating data in previously unseen scenarios.

CCS Concepts

• **Information systems** → **Spatial-temporal systems**.

*Corresponding author



KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737039

Keywords

Human Mobility; Trajectory Generation; Generative Model; Mobility Simulation; Diffusion Model

ACM Reference Format:
Bangchao Deng, Ling Ding, Lianhua Ji, Chunhua Chen, Xin Jing, Bingqing Qu, and Dingqi Yang. 2025. Marionette: Fine-Grained Conditional Generative Modeling of Spatiotemporal Human Trajectory Data Beyond Imitation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3711896.3737039

KDD Availability Link:
The source code of this paper has been made publicly available at https://doi.org/10.5281/zenodo.15523812.

1 Introduction

Human trajectory data analytics offer valuable insights into individual movement and collective mobility patterns, which can support a wide range of applications, including urban planning [62], traffic control [35], and crowd monitoring [5], etc. However, the availability of publicly accessible and high-quality trajectory datasets is becoming severely limited due to privacy concerns and stringent regulatory requirements [21]. In response to these challenges, synthetic human trajectory data emerges as an alternative solution by generating artificial trajectories that resemble real human trajectories under certain spatiotemporal and statistical similarities without leaking any sensitive information.

Toward this goal, early approaches captured the regularity of human mobility behaviors by designing mechanistic models with explicit physical meanings under a small number of parameters [15, 42]. However, these models, mostly based on simple assumptions on human mobility (e.g., exploration and preferential return model [41]), often fail to accurately capture the complex mobility patterns encoded by real-world human trajectories [8]. To address this limitation, recent works resort to neural network-based generative models for learning directly from real-world human trajectory data to imitate human mobility behaviors, using Recurrent Neural

Networks (RNNs), attention mechanisms [47], Generative Adversarial Networks (GANs) [3, 12, 56], Variational Autoencoders (VAEs) [22, 33], or diffusion models [57, 63, 64]. Although the expressiveness of these neural architectures largely improves the capability of these models to capture the complex spatiotemporal mobility patterns, such an unconditional generative paradigm relies on a pure data-driven learning scheme without considering the rich context of human mobility, making these models inflexible and uncontrollable to meet the requirements of user-specified scenarios (e.g., generating trajectory data in previously unseen scenarios).

Existing studies have demonstrated that human mobility patterns are profoundly influenced by various context factors, such as user profile [39], almanac [5], weather conditions [4], social events [20], festive activities [24], or other urban anomalies [30]. For example, user profiles often determine user preference for locations in the user trajectories [39]; daily mobility patterns have been shown to vary across different days in a week [9, 53], working days versus holidays, and the days of social events such as cultural festivals or concerts [20, 24]; moreover, empirical evidence also shows that both routine weather variations and extreme climatic events (e.g., rainstorms or typhoons) can significantly alter the mobility patterns of populations [4, 7]. Subsequently, a few recent works on human trajectory generation started to consider the context information, mostly user profiles [8, 33], or spatiotemporal constraints [32], to condition the generated trajectory data. However, these approaches are designed to leverage the user profile as a global context to condition the entire user trajectory as a whole, while overlooking the dynamics of some context factors where a partial context may influence part of the trajectory only. For example, a rainstorm often lasts for a few hours, which influences the mobility patterns of human daily trajectories over its active period only. Compared to the global context which provides coarse-grained conditioning to trajectory generative models, the partial context allows fine-grained conditioning in a more flexible and controllable manner.

Against this background, we study the problem of fine-grained conditional generative modeling of spatiotemporal human trajectory data. Specifically, different from the periodic and regularly sampled vehicle trajectories with GPS coordinates such as taxi trajectories [3], human trajectories usually consist of sparse and irregularly observed presence events on continuous time at semantic-enriched locations, such as user voluntarily shared check-ins at Points of Interest (POIs) on social media [10, 51]. Subsequently, those widely studied (conditional) sequence models for language modeling, which model only the order of tokens in sequence, are incapable of capturing important temporal information over continuous time. In this context, as a human trajectory consisting of a sequence of stochastic presence events over continuous time, a few recent works resort to neural Temporal Point Processes (TPPs) to model human trajectory data [8, 33, 57, 58], showing superior performance. Following the generative process of a TPP, these models generate a sequence of user presence events in an autoregressive manner where the intensity function of the TPP is modeled by a neural network parameterized by information extracted from historical events. This scheme can easily incorporate the global context by adding it as an initial (global) condition of the TPP, which influences the intensity function over the whole generation process [8, 33, 57, 58]. However, it is challenging to integrate the partial context during the autoregressive generation process of the TPP, because the partial context influences the intensity function only during its active period. It is also impractical to design entry-exit mechanisms for the partial context influencing the intensity function during the autoregressive generation process; because the intensity function can only shift when an event occurs, which does not necessarily align with the starting and ending time of the partial context (e.g., a user presence event is rarely observed right at the starting and ending time of a rainstorm).

To address these issues, we propose Marionette, a **Ma**nipulatable gene**r**at**i**ve m**o**del for human trajectory data with fine-grained conditions. First, we integrate the mobility-related global and partial contexts using a context alignment method, and then extract both sequence-level and event-level conditions. Second, for conditional temporal modeling, we design a conditional TPP following a diffusion-alike TPP scheme to capture the impact of the conditions on both event-triggered shifts and continuous variation of the intensity function of the TPP. Third, for conditional spatial modeling, we design a cascading conditioning mechanism integrated with a discrete diffusion model to capture the cascading dependence of locations on activity categories and further on time. Our Marionette can thus offer fine-grained controllable generative modeling of human trajectory data. Our contributions are summarized below:

- We reveal the limitations of existing human trajectory generative models in ignoring the significant impact of rich context information on human mobility patterns, which limits the flexibility and manipulability in data-driven mobility simulation.
- We propose Marionette, a **Ma**nipulatable gene**r**at**i**ve m**o**del for human trajectory data with fine-grained conditions, offering fine-grained controllable generative modeling of human trajectory data with both global and partial mobility-related contexts.
- We conduct a thorough evaluation of Marionette on two real-world human trajectory datasets against a sizeable collection of state-of-the-art baselines. Results show that Marionette consistently outperforms the best baselines by 13.96-54.13% on statistical and distributional similarity metrics and by 9.36-40.63% in task-based data utility evaluation. Ablation studies verify our key design choices. Case studies also demonstrate the manipulability of Marionette in generating data in previously unseen scenarios.

2 Related Work

2.1 Context Modeling for Mobility

Existing studies have widely shown that human mobility patterns are complicated and influenced by various contextual factors, such as weather conditions [7, 20], social events [55], and government policies [14]. For example, Chapman et al. [4] revealed that the mobility patterns of the urban population usually exhibit significant variations in different temperatures; Chen et al. [7] investigated the impact of extreme weather events (such as typhoons) on urban human flow and found that destructive weather events led to significant changes in the local spatial agglomeration of migrant populations; Jiang et al. [24] predicted crowd dynamics under different event situations (e.g., earthquakes, typhoons and national festivals); Gao et al. [14] studied the social distancing policies of different regional governments in the United States and their impact on human mobility patterns during the COVID-19 period, and revealed

that the implementation of different policies under the pandemic had a significant impact on human mobility; STORM [20] predicts abnormal crowd traffic by modeling the time-occasional impact of sudden event context (e.g., weather conditions); CausalMob [55] incorporates Large Language models (LLMs) and causal framework to learn confounder representations between social events and human mobility for prediction; CDGON [30] integrates existing physical knowledge and the neural ODE model to capture the dynamic recovery patterns of urban mobility in post-disaster scenarios. However, all these existing techniques focus on the predictive modeling of human mobility, which differs from our focus on the generative modeling for human trajectories with fine-grained conditions of both global and partial contexts.

2.2 Trajectory Generative Modeling

Existing works on trajectory generative modeling can be broadly classified into two categories as follows.

First, *information-free methods* focus on directly learning human mobility patterns from data. Earlier approaches mostly designed mechanistic models with a small number of parameters to describe the key characteristics of human mobility [15, 42], such as temporal periodicity and spatial continuity. For example, the Exploration and Preferential Return (EPR) model [41] integrates exploration and return mobility patterns by selecting new locations through a random walk process for exploration, and revisiting locations based on their visit frequency for preferential return. TimeGeo [25] expands the EPR model by incorporating temporal elements to better characterize temporal mobility patterns. With advancements in deep learning, recent methods make fewer prior assumptions, allowing the model to capture more complex spatiotemporal patterns inherent in real-world mobility trajectories. For instance, SeqGAN [56] uses Generative Adversarial Networks (GANs) for sequence generation. MoveSim [12] introduces physical distance, temporal periodicity, and historical transition matrices into a GAN framework. TrajGen [3] uses a CNN-based GAN to map mobility trajectories to images and generate synthetic trajectory images, which are then processed by a Seq2Seq model to produce the synthetic trajectories. SAVE [23] integrates VAE and LSTM to generate mobility trajectories. COLA [47] develops a model-agnostic transfer framework by separating private and shared modules and calibrating prediction probabilities to account for city-specific characteristics in human trajectory simulation. Recently, neural temporal point processes [6, 40, 57, 65] are widely used to model the stochasticity of temporal dynamics of sparse and irregularly observed human trajectories. In the context of trajectory generation, VOLUNTEER [33] combines a two-layer VAE model with a temporal point process to capture human mobility characteristics. ActSTD [58] improves dynamic modeling of individual trajectories by using neural ordinary differential equations in the continuous location domain. DSTPP [57] further models complex spatiotemporal joint distributions using diffusion models. Geo-CETRA [32] decomposes the spatiotemporal constraints of human trajectories and models time and spatial distributions separately, generating high-quality trajectories that satisfy constraints under a beam search strategy. MIRAGE [8] integrates an intensity-free neural TPP and a neural EPR model to imitate the human decision-making process in trajectory generation. Despite their advancements, these methods often fail to consider the impact of various mobility-related contexts on human mobility, making them inflexible and uncontrollable to meet the requirements of user-specified generation scenarios.

Second, *information-guided methods* incorporate additional contexts, such as road networks, origin-destination pairs, and other relevant factors, into the mobility generative models, to further regulate the generated human mobility patterns. For instance, TS-TrajGen [26] combines the A* algorithm [17] with a GAN framework to generate continuously sampled trajectories on given urban road networks. DiffTraj [63] applies a diffusion probabilistic model guided by external factors, such as departure times and regions, to generate trajectories in the continuous spatial domain. ControlTraj [64] pre-trains a road network encoder and employs a conditional diffusion model under the structural constraints of the road network topology. However, these works all focus on continuously sampled vehicle trajectory data (i.e., regularly sampled GPS traces) and subsequently provide certain controllable generation in the spatial domain (e.g., generating GPS points on a given road network), which differs from our current work focusing on sparse and irregularly observed human trajectories at semantic-enriched locations and its controllable conditional generative modeling with fine-grained contexts.

3 Problem Formulation

We present our key terminology and problem formulation below.

Human Trajectory. A human trajectory is defined as a time-ordered sequence $\psi = \{(t_1, k_1, l_1), (t_2, k_2, l_2), ..., (t_i, k_i, l_i), ...\}$, where each tuple (t_i, k_i, l_i) is a presence event defined as a tuple consisting of a timestamp t_i, and a semantic activity category k_i, and a location (POI[1]) l_i.

Human Mobility-Related Context. Mobility-related contexts refer to environmental factors that may have an impact on human mobility patterns, such as almanac [5], weather conditions [4], or social events [20], etc.

Conditional Trajectory Generation. Given a real-world human trajectory dataset and various human mobility-related conditions, the objective is to generate a synthetic trajectory dataset according to user-specified mobility-related conditions while preserving the fidelity and utility of the original real-world dataset.

4 Marionette

Figure 1 shows the overview of our Marionette consisting of three parts: 1) context alignment and extraction, 2) conditional temporal modeling, and 3) conditional spatial modeling. First, for a given trajectory, the context alignment and extraction module integrates the mobility-related global and partial contexts using a context alignment method, and then extracts both sequence-level (for the whole sequence) and event-level (for each presence event) conditions. Second, the conditional temporal model designs a conditional Temporal Point Process (TPP) on top of a diffusion-alike TPP model to capture the impact of the conditions on both event-triggered shifts and continuous variation of the intensity function of the

[1]A POI is a semantic-enriched location, such as a restaurant or a shop; in the context of human trajectories, we do not distinguish the two terms throughout this paper.

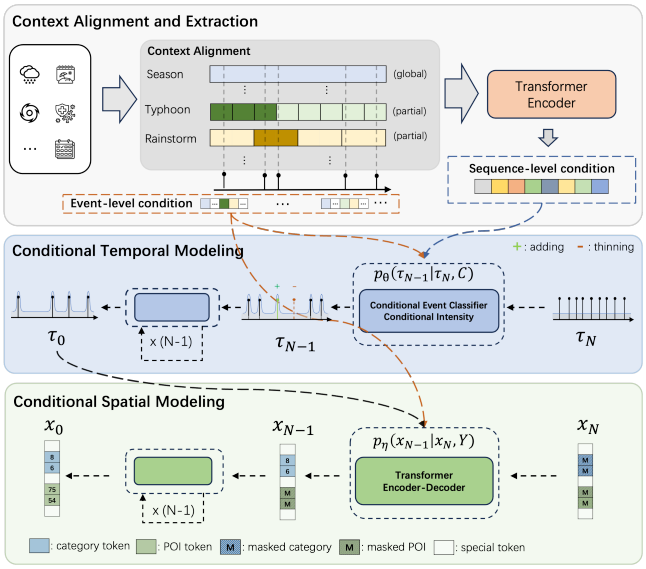


Figure 1: Overview of Marionette.

TPP. Third, the conditional spatial model designs a cascading conditioning mechanism integrated with a discrete diffusion model to capture the dependence of locations on activity categories and further on time.

4.1 Context Alignment and Extraction

The context of human mobility is rather complex, and its impact may persist for varying durations. In this work, we define two types of mobility-related context according to the impact duration. First, a global context influences the entire trajectory, serving as a coarse-grained condition for the whole sequence of the trajectory. For example, considering generating daily trajectories, the weekday or seasonal information is persistent over the whole trajectory. Second, a partial context influences part of the trajectory only, serving as a fine-grained condition for a subsequence of the trajectory. For example, a rainstorm usually lasts for several hours, thus having an impact on the mobility patterns of user trajectories over its period only. Moreover, this impact on the mobility patterns further varies across the absolute time of the rainstorm [5]; for instance, a significant impact is usually observed if the rainstorm happens on a weekend morning (as people may avoid unnecessary travel), while the impact is often negligible if it happens over the night.

4.1.1 ***Context Alignment.*** To systematically integrate the global and partial contexts while minimizing information redundancy or loss, we propose a mobility context alignment method, as shown in Figure 1. Specifically, as the partial context impacts only a subperiod of the whole trajectory, we first segment the whole duration of the trajectory into a number of L segments, where the length of each segment corresponds to the finest granularity of all partial contexts. For example, rainstorm signals are reported at an hourly granularity, while typhoon signals are reported at a half-hour granularity; for a daily trajectory of 24 hours, the partial context of these weather conditions then falls into $L = 48$ segments. Subsequently, a partial context for the trajectory is represented by a vector $c_p \in \mathbb{R}^L$. Afterward, as the global context persists over the whole trajectory, we extend the global context to match the length of the partial context vector represented as $c_g \in \mathbb{R}^L$, obtaining aligned global and partial contexts for each time segment. Finally, we concatenate all extended global contexts and partial contexts along with the time segment into a sequence of context $C \in \mathbb{R}^{|C|\times L}$, where $|C|$ represents the number of contexts.

The advantage of this alignment mechanism lies in its flexibility to integrate contexts of different time granularities with trajectories of different lengths. For example, for a daily trajectory, the day-in-week (e.g., Monday) is considered as a global context because it is persistent for the whole day; for a weekly trajectory, the day-in-week becomes a partial context because it is valid only on the first day of the week (Monday).

4.1.2 ***Context Extraction.*** With the aligned global and partial contexts, we then extract two levels of conditions, i.e., event-level and sequence-level conditions. First, for each presence event i in continuous time, we get its mobility-related context based on its time t_i by probing into the sequence of aligned context C, as shown in Figure 1; the obtained contexts are then embedded into a latent space as learnable embeddings to obtain the event-level condition, denoted as $e_i^c \in \mathbb{R}^{|C|\times D}$, where i represents the i-th event, $|C|$ represents the number of contexts, and D is the dimension of the embedding space. *The event-level condition allows us to capture the fine-grained condition at the exact presence event.* Second, we also consider the context of the entire sequence. We embedded the aligned global and partial contexts $C \in \mathbb{R}^{|C|\times L}$ into a latent space via a transformer encoder, to output the sequence-level condition, denoted as $C_{seq} \in \mathbb{R}^{|C|\times L\times D}$, where D again represents the dimension of the embedding space. Note that the sequence-level condition integrates the contexts across all time segments over the whole trajectory duration *regardless of when an event occurs.* In other words, different trajectories (consisting of different presence events) may have the same sequence-level contexts. Therefore, *the sequence-level condition captures the temporal dynamics and interactions of different contexts over the whole trajectory.*

4.2 Conditional Temporal Modeling

As a human trajectory intrinsically consists of a sequence of stochastic presence events over continuous time, neural TPPs have been recently adopted to model the stochasticity of presence events in the sequence. Specifically, existing neural TPPs model a sequence of presence events in an autoregressive manner where the intensity function of the TPP is modeled by a neural network parameterized by information extracted from historical events [8, 33, 57, 58]. However, such an autoregressive scheme is incompatible with the partial context that is active on a subsequence only; it is also impractical to design an entry-exit mechanism for the partial context influencing the intensity function during the autoregressive generation process; because the intensity function can only shift when an event occurs, which does not necessarily align with the starting and ending time of the partial context as shown in Figure 1 (e.g., no user presence event is observed right at the starting and ending time of a rainstorm). Against this background, instead of adopting an autoregressive scheme, we resort to a diffusion-alike TPP

scheme and design a conditional TPP capturing the impact of the conditions on both event-triggered shifts and continuous variation of the intensity function of the TPP.

4.2.1 ***Diffusion-alike TPP.*** The diffusion-alike TPP [34] considers the forward process of the diffusion model as operations involving removing old events from the current sequence and adding new events from a homogeneous Poisson process. Let $\tau_0 = (t_1, ..., t_j, ...)$ denote a sequence sampled from a TPP specified by an unknown intensity λ_0. The forward process can be defined as a sequence of TPPs that start with the true intensity λ_0 and converge to a homogeneous Poisson process (HPP), i.e., $\lambda_0 \to \lambda_1 \to ... \to \lambda_N$:

$$\lambda_n(\tau) = \alpha_n \lambda_{n-1}(\tau) + (1 - \alpha_n)\lambda_{HPP} \tag{1}$$

where $1 > \alpha_1 > \alpha_2 > ... > \alpha_N > 0$ and λ_{HPP} denotes the constant intensity of an HPP. Eq. 1 indeed describes a combination of two processes: 1) a process λ_{n-1} thinned with probability $1-\alpha_n$ (removing old events), and 2) an HPP with intensity $(1-\alpha_n)\lambda_{HPP}$ (adding new events). Importantly, due to the property of the Markov chain, we can derive the intensity of $\lambda_n(\tau)$ at arbitrary timestep in the diffusion process directly from $\lambda_0(\tau)$ as:

$$\lambda_n(\tau) = \overline{\alpha}_n \lambda_0(\tau) + (1 - \overline{\alpha}_n)\lambda_{HPP} \tag{2}$$

The reverse process is designed to sample realizations $\tau_0 \sim \lambda_0$ starting from the noise-corrupted $\tau_N \sim \lambda_{HPP}$ by learning to reverse the Markov chain of the forward process, i.e., $\lambda_N \to ... \to \lambda_0$. The events in the sequence τ_{n-1} are decomposed into disjoint sets of events based on whether they are also in τ_0 or τ_n, which can be generated by sampling from the intensity functions $\lambda_{n-1}(\tau|\tau_0, \tau_n)$. Since there is no information about τ_0 in the reverse process, it is learned to be approximated as $\hat{\tau}_0 \approx \tau_0$ below:

$$\lambda_\theta(\tau) = H \sum_{m=1}^{M} \omega_m f(\tau; \mu_m, \sigma_m) \tag{3}$$

$$g_\theta = MLP(e) \tag{4}$$

Eq. 3 includes the unnormalized mixture of M weighted and truncated Gaussian density functions to parameterize the events $\tau_0 \backslash \tau_n$ (events belong to τ_0 but not belong to τ_n), and H represent the number of events in τ_n. Eq. 4 represents a classifier of the events $\tau_0 \cap \tau_n$ (events from τ_n belong to τ_0, where e represents a learnable event time embedding (more details below). By learning $\lambda_\theta(\tau)$ and g_θ, we can approximate $\hat{\tau}_0 \approx \tau_0$ and then use the posterior intensity to reverse the noising process. Please refer to [34] for more details on the diffusion-alike TPP.

In the following, we design a conditioning mechanism for this reverse process to offer more flexible and controllable generation. We present our conditional design of g_θ and $\lambda_\theta(\tau)$ below.

4.2.2 ***Conditional Event Classifier g_θ.*** To condition the classifier g_θ, we first encode each event time $t_i \in \tau_n$ and inter-event time $\Delta_i = t_i - t_{i-1}$, to produce an event time embedding e_i^t using a sinusoidal embedding [45] which characterize the time information of the event, and then integrates it with its corresponding event-level condition e_i^c (obtained in Section 4.1.2) to consider the rich context of the event. We feed the concatenated e_i^t and e_i^c into the event classifier g_θ parameterized by an MLP:

$$g_\theta = MLP([n, e_i^t; e_i^c]) \tag{5}$$

where n refers to the diffusion timestep. The conditional g_θ thus captures the event-triggered shifts over a trajectory sequence.

4.2.3 ***Conditional Intensity $\lambda_\theta(\tau)$.*** To condition the intensity function $\lambda_\theta(\tau)$, we parameterize the Gaussian mixture in $\lambda_\theta(\tau)$. Specifically, different from the autoregressive TPPs where the conditional intensity characterizes the intensity function at a given time conditioned on the past event history, in the diffusion-alike TPP, the intensity function characterizes temporal dynamics over the whole sequence. Subsequently, instead of using the past events only, we can now use all events in the sequence as conditions to capture the continuous variation of the intensity function over the whole sequence. To this end, we first encode all event time information by feeding all e_i^t where $t_i \in \tau_n$ to a neural network to output a sequence embedding $\bar{s}$ which characterizes the overall temporal dynamics of events in the trajectory, and then aggregate it with the sequence-level context C_{seq} (obtained in Section 4.1.2) to consider the rich context of the whole trajectory. Afterward, we use the concatenated $\bar{s}$ and C_{seq} as inputs to compute the Gaussian mixture parameters $\omega_m, \mu_m, \sigma_m$ of $\lambda_\theta(\tau)$ in Eq. 3 via an MLP as follows:

$$\begin{cases} \omega_m = \text{Softplus}(MLP_\omega([n, \bar{s}, C_{seq}])) \\ \mu_m = \text{Sigmoid}(MLP_\mu([n, \bar{s}, C_{seq}])) \\ \sigma_m = \exp(-|MLP_\sigma([n, \bar{s}, C_{seq}])|) \end{cases} \tag{6}$$

where n refers to the timestep of the diffusion process; the Softplus, Sigmoid, and exponential (exp) functions are used to put the constraints on the distribution parameters ($\omega_m > 0$, $0 < \mu_m < 1$) and regularization on $\sigma_m \to 0$).

In summary, the training process of the conditional temporal model learns the conditional event classifier and the conditional intensity function by minimizing the sum of the respective Binary Cross-Entropy (BCE) loss and Negative Log-Likelihood (NLL) loss, i.e., $\mathcal{L}_{temporal} = \mathcal{L}_{NLL} + \mathcal{L}_{BCE}$.

4.3 Conditional Spatial Modeling

Different from the continuously sampled vehicle trajectories with GPS coordinates such as taxi trajectories [3], the human trajectories involve the semantic-rich locations (POIs) in the spatial domain. Subsequently, the conditional spatial model is designed to generate the discrete POIs. Recent works have shown that the human decision-making process of generating trajectory follows a cascading manner [8, 58]. At a given time (e.g. noon), an activity category (e.g., food) is first sampled according to the time; a POI (e.g., a specific restaurant) is then sampled according to the time and activity category. Therefore, we design a cascading conditioning mechanism integrated with a discrete diffusion model to capture the dependence of locations on activity categories and further on time.

We first integrate the activity category and POI information into the same sequence by introducing three special tokens <sos>, <eos>, and <sep>, represented as:

$$\text{x} = (\texttt{<sos>}\, k_1, k_2, \ldots, k_j\, \texttt{<sep>}\, l_1, l_2, \ldots, l_j\, \texttt{<eos>}) \tag{7}$$

where k_j and l_j represent category tokens and POI tokens, respectively. The <sos> and <eos> tokens denote the start and end of a sequence, respectively; the <sep> token serves as a separator between the category sequence and the POI sequence.

Based on this representation, we then employ a discrete diffusion model [1] and design a cascading conditioning mechanism on top of it to model the conditional spatial dynamics of human trajectories. Specifically, given a sequence $x_0 \sim q(x_0)$, the forward process gradually corrupts it via a fixed Markov chain, yielding a sequence of increasingly noisy latent variables $x_{1:N} = x_1, x_2, ..., x_N$: $q(x_{1:N}|x_0) = \prod_{n=1}^{N} q(x_n|x_{n-1})$. In the reverse process, we gradually denoise the noisy sequence x_N to recover the sequence x_0 conditioned on an encoded condition Y (more details below in Section 4.3.2):

$$p_\eta(x_{0:N}) = P(x_N) \prod_{n=1}^{N} p_\eta(x_{n-1}|x_n, Y) \tag{8}$$

Moreover, as our spatial sequence now contains three types of discrete tokens, i.e., activity category tokens, POI tokens, and special tokens, we design a customized transition forward process with justified design choices as follows.

4.3.1 ***Block-wise Transition Forward Process.*** We adopt a block-wise transition matrix [59] that prevents internal transitions across different token types during the forward process, thereby reducing the transition matrix Q_n to a block-wise diagonal structure.

$$Q_n = \begin{bmatrix} Q_n^{spec} & & \\ & Q_n^{cat} & \\ & & Q_n^{poi} \end{bmatrix} \tag{9}$$

where Q_n^{spec}, Q_n^{cat}, and Q_n^{poi} represent the probabilities of the internal transition within special tokens, category tokens, and poi tokens, respectively.

First, as special tokens describe the structure of the sequence, any transition between them will lead to an invalid sequence. Therefore, we choose to disable any transition between them by setting $Q_n^{spec} = \mathrm{I}$.

Second, since category tokens represent the semantic information of locations, transitioning a category token to another category token would cause abrupt changes in the semantic meaning of locations. To address this, we choose to transition a category token to a special [*MASK_CAT*] token rather than to another meaningful category token, where [*MASK_CAT*] retains its own state without further transitions. Therefore, we introduce the absorbing state transition matrix for category tokens:

$$Q_n^{cat} = \begin{bmatrix} 1-\gamma_n^{cat} & 0 & \cdots & 0 \\ 0 & 1-\gamma_n^{cat} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_n^{cat} & \gamma_n^{cat} & \cdots & 1 \end{bmatrix} \tag{10}$$

Third, transitions between POI tokens do not cause significant changes in the overall semantic meaning of the sequence, while also enhancing the diversity of generated trajectories. Therefore, we adopt the Mask-and-replace diffusion strategy [16]: each token has a probability of γ_n^{poi} to be replaced by the [*MASK_POI*] token, a probability of $K\beta_n$ to transit to any other POI token uniformly, and a probability of $\alpha_n = 1 - K\beta_n - \gamma_n^{poi}$ to remain unchanged. Notably, the [*MASK_POI*] token always retains its own state:

$$Q_n^{poi} = \begin{bmatrix} \alpha_n+\beta_n & \beta_n & \beta_n & \cdots & 0 \\ \beta_n & \alpha_n+\beta_n & \beta_n & \cdots & 0 \\ \beta_n & \beta_n & \alpha_n+\beta_n & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_n^{poi} & \gamma_n^{poi} & \gamma_n^{poi} & \cdots & 1 \end{bmatrix} \tag{11}$$

Furthermore, as the sample of POIs depends on the sample of activity categories, we use different rates for transforming category and POI tokens into their final noisy states (i.e., categories transition to [*MASK_CAT*] and POIs transition to [*MASK_POI*]). Specifically, the masking rate for categories follows a non-linear schedule, where it is initially low and then rapidly increases to transform categories into [*MASK_CAT*] tokens in the later stages, while POIs follow a conventional linear noising schedule [1]. *This design ensures that during the reverse process, category tokens are generated first from noisy tokens, naturally guiding the generation of POIs.*

Through the above operations, the sequence x_0 in Eq. 7 (containing the POI tokens and their corresponding category tokens, as well as special tokens) is corrupted via the transition probability matrix, which gradually transfers all category and POI tokens into their respective masked tokens x_N with no meaningful information after N steps of state transitions. Next, we will introduce the reparameterization of the reverse process conditioned on the context.

4.3.2 ***Cascading Conditioning Denoising.*** In the reverse process, the initial x_N is denoised iteratively by neural networks conditioned on both time and context embeddings to obtain x_0. Our cascading conditioning mechanism leverages the generated event time along with the corresponding mobility-related condition to guide the denoising process of the discrete diffusion model. To condition the p_η, we first extract the event-level condition e_i^c for each check-in based on the event time. Afterward, we input the entire sequence of conditions $(e_1^c, e_2^c, \ldots, e_j^c, ...)$ into a transformer encoder to out an encoded condition Y. We then input it along with the current step n of the diffusion process into the Transformer decoder to generate an entire sequence. Following [59, 61], here we adopt the x_0-parameterization approach that $p_\eta(x_{n-1}|x_n, Y) \propto \sum_{\tilde{x}_0} q(x_{n-1}, x_n|\tilde{x}_0)\, p_\eta(\tilde{x}_0|x_n, Y)$. Based on the reparameterization trick, our denoising objective encourages the network to predict a noiseless sequence with the following loss function: $\mathcal{L}_{spatial} = -\log p_\eta(x_0|x_n, Y)$

4.4 Model Training and Generation

The training process of our Marionette minimizes the sum of temporal and spatial losses:

$$\mathcal{L} = \mathcal{L}_{temporal} + \mathcal{L}_{spatial} \tag{12}$$

The trajectory generation process takes the following steps. For user-specified global and partial contexts, we first perform the context alignment and directly get the sequence-level condition. Subsequently, we generate a sequence of event times using the conditional TPPs, where the event-level conditions are obtained on the fly during the diffusion reverse process by probing the event time into the sequence of aligned context C, as shown in Figure 1.

Table 1: Dataset statistics

	Campus	Istanbul
#Trajectories	31,819	11,949
#Presence events	628.7K	74.1K
#POIs	396	3465
#Categories	8	9
Sampling Rate	27s (median)	3.60h (median)
Activity Scale	Campus	City
Duration	24 months	22 months

Finally, based on the obtained sequence of event times and event-level conditions, we generate the corresponding activity categories and POIs and finally obtain a complete trajectory.

5 Experiments

5.1 Experimental Settings

5.1.1 Dataset. We conduct extensive experiments on two human trajectory datasets: a **Campus** human trajectory dataset based on Wi-Fi connection records [5], and a **Istanbul** dataset collected from a location-based social network Foursquare [53, 54]. Table 1 shows the dataset statistics. For the campus dataset, we consider the daily trajectory with the global context including Day-in-weeks, Almanac, Teaching Arrangement, and the partial context including Hour-in-day and Weather Conditions (Tropical Cyclone Warning Signal and Rainstorm Warning Signal). For the Istanbul dataset, we consider the daily trajectory with the global context including Day-in-weeks, Social Events (Ramadan), Public holidays, and the partial context including Hour-in-day.

5.1.2 Baselines. We consider the following state-of-the-art baselines of three categories: statistical models **Semi-Markov** [28] and **TimeGeo** [25]; neural TPPs **RMTPP** [11], **ERTPP** [49], and **ActSTD** [58]; deep learning generative models **SeqGAN** [56], **MoveSim** [12], **VOLUNTEER** [33], **DiffTraj** [63], **COLA** [47], and **MIRAGE** [8]. We also extend MIRAGE by feeding the same contexts we used as an additional input to its autoregressive TPP, denoted as **Cond-MIRAGE**, to compare with our method that uses a diffusion-alike TPP. The baseline details are in Appendix A.

5.1.3 Statistical and Distributional Similarity Metrics. We adopt five popular metrics [12, 36] to evaluate the resemblance between real and generated trajectories in different aspects. **Distance** measures the distance between successive locations in a trajectory. **Radius** of gyration is calculated as the root mean squared distance of all locations from the central one in a trajectory. **Interval** is computed as time intervals between successive events in a trajectory. **DailyLoc** computes the unique locations visited by users. **Category** computes the overall distribution of the POI categories. **G-RANK** computes the visited frequency of the top locations. We use the Jensen-Shannon divergence (JSD) [13] as the similarity metric between the distributions of real and generated trajectories.

5.1.4 Task-Based Utility Evaluation Protocol. We adopt the comprehensive task-based evaluation protocol[2] [8] to evaluate the utility of the generated trajectories in four downstream predictive mobility modeling tasks, i.e., location recommendation (LocRec), next location prediction (NexLoc), semantic location labeling (SemLoc), and epidemic simulation (EpiSim), which model user trajectory data in four different aspects, i.e., user preferences, sequential patterns, collective traffics, and spatiotemporal contacts, respectively. For each task, multiple state-of-the-art techniques are used to conduct experiments and report the results on multiple metrics. Finally, this protocol measures the paired performance discrepancy between the real and generated trajectories using Mean Absolute Percentage Error (**MAPE**) and Mean Squared Percentage Error (**MSPE**). A smaller performance discrepancy here indicates better generative modeling performance; because the generated trajectories show a more similar utility to the real trajectories across these mobility modeling tasks. Please refer to Appendix B for more details.

[2]https://github.com/UM-Data-Intelligence-Lab/MIRAGE

5.2 Performance Comparison

Table 2 shows the performance on statistical and distributional similarity metrics, and Table 3 shows the performance in the task-based evaluation on MAPE (we observe similar results on MSPE, shown in Appendix C). We observe that Marionette achieves the best performance in most cases. Specifically, on statistical and distributional similarity metrics, Marionette yields 54.13% and 13.96% improvement (reduction on JSD) over the best-performing baselines on the Campus and Istanbul datasets, respectively. Meanwhile, in task-based evaluation, Marionette achieves 40.63% and 9.36% improvement over the best baselines on the two respective datasets. The large improvement on the Campus dataset is attributed to its richer contexts compared to the Istanbul dataset.

We also have interesting findings on our extended conditional baseline Cond-MIRAGE, where we encode both our global and partial contexts (the same contexts as for our Marionette) into an overall condition and feed it as an additional input to further parameterize the autoregressive TPP of MIRAGE. On one hand, Cond-MIRAGE outperforms its original version MIRAGE, showing the usefulness of considering the mobility-related context in trajectory generative models. On the other hand, our Marionette further outperforms Cond-MIRAGE using the same contexts, which is attributed to the superiority of our model design. In other words, encoding the global and partial contexts as an overall condition fails to offer fine-grained conditioning from the partial contexts, thus leading to inferior performance.

5.3 Ablation Study

We conduct an ablation study with the following two variants, i.e., **Marionette-noCond** which is a variant of Marionette trained without any context and **Marionette-noPartial** which is a variant of Marionette trained without the partial context, on the Campus dataset due to its rich contexts. Tables 4 and 5 show the results. We observe that Marionette consistently outperforms Marionette-noCond on both similarities and task-based evaluation by 32.96% and 30.06%, respectively, highlighting the importance of contexts in trajectory generative modeling. Second, Marionette also consistently outperforms Marionette-noPartial with an improvement of 37.22% and 23.87% on similarities and task-based evaluation, respectively, which verifies the usefulness of the partial context.

Table 2: Performance on Statistical and Distributional Similarity Metrics

	Campus						Istanbul					
Method	Distance	Radius	Interval	DailyLoc	Category	G-RANK	Distance	Radius	Interval	DailyLoc	Category	G-RANK
Semi-Markov	0.2639	0.2209	0.1240	0.1342	0.0433	0.1214	0.4796	0.3190	0.1238	0.1434	0.0343	0.2433
Time Geo	0.2940	0.2827	0.1675	0.1022	0.0416	0.1597	0.5463	0.5536	0.0724	0.1077	0.0442	0.2962
RMTPP	0.2508	0.2270	0.0836	0.0799	0.0506	0.1254	0.1881	0.1449	0.0837	0.1406	0.0364	0.2647
ERTPP	0.1683	0.1269	0.0571	0.0365	0.0356	0.1007	0.2763	0.2126	0.0962	0.1390	0.0782	0.2202
ActSTD	0.1581	0.1901	0.0296	0.0836	0.0303	0.1546	0.3493	0.1913	0.0394	0.0857	0.0610	0.1849
SeqGAN	0.1844	0.1487	0.0704	0.0676	0.0311	0.1407	0.1782	0.1638	0.1375	0.0874	0.0265	0.1989
MoveSim	0.1931	0.1256	0.1095	0.1231	0.0464	0.2202	0.1676	0.1576	0.0471	0.1562	0.0525	0.1879
VOLUNTEER	0.1823	0.1971	0.0735	0.0957	0.0464	0.1542	0.1994	0.1657	0.0299	0.0766	0.0343	0.1922
DiffTraj	0.1520	0.1214	0.1404	0.0855	0.0397	0.1064	0.1974	0.1676	0.0920	0.0773	0.0248	0.2497
COLA	0.1957	0.2062	0.2869	0.1554	0.0312	0.1016	0.1542	0.1538	0.0998	0.1469	0.0311	0.1349
MIRAGE	0.1250	0.0656	0.0135	0.0428	0.0252	0.0829	0.1282	0.0559	0.0123	0.0411	0.0061	0.1288
Cond-MIRAGE	0.0937	0.0635	0.0128	0.0238	0.0157	0.0580	0.1130	0.0698	0.0094	0.0392	0.0054	**0.1172**
Marionette	**0.0664**	**0.0134**	**0.0083**	**0.0144**	**0.0024**	**0.0248**	**0.1060**	**0.0450**	**0.0080**	**0.0384**	**0.0029**	0.1238

Table 3: Performance in the Task-Based Evaluation on MAPE

	Campus				Istanbul			
Method	LocRec	NexLoc	SemLoc	EpiSim	LocRec	NexLoc	SemLoc	EpiSim
Semi-Markov	0.7501	0.6630	0.3738	0.7478	0.9830	0.9630	0.3632	0.6237
Time Geo	0.6667	0.7879	0.2977	0.4506	0.8490	0.9476	0.3104	0.7781
RMTPP	0.5607	0.5707	0.3180	0.1966	0.6231	0.5746	0.4640	0.4112
ERTPP	0.4258	0.6023	0.3103	0.1389	0.7436	0.6323	0.3490	0.2914
ActSTD	0.3781	0.5692	0.6553	0.7896	0.7774	0.7249	0.3438	0.5142
SeqGAN	0.5185	0.8488	0.3669	0.8125	1.0636	0.7816	0.2728	0.4743
MoveSim	1.8991	0.5324	0.4538	0.1595	0.7316	0.7927	0.3363	0.2623
VOLUNTEER	0.5712	0.7534	0.6100	0.7245	0.9562	0.8908	0.3886	0.1917
DiffTraj	0.9795	0.4508	0.2645	0.5634	0.8418	0.7759	0.3713	0.4574
COLA	0.6891	0.4449	0.4645	0.1596	0.6015	0.7193	0.3953	0.2921
MIRAGE	0.3309	0.3856	0.1273	0.1180	0.5628	0.5002	0.1750	0.1434
Cond-MIRAGE	0.3018	0.2128	0.1149	0.0653	0.3942	0.3567	**0.1429**	0.1640
Marionette	**0.1862**	**0.0847**	**0.0943**	**0.0352**	**0.3303**	**0.3312**	0.1564	**0.1096**

Table 4: Ablation Study on Statistical and Distributional Similarity Metrics

Method	Distance	Radius	Interval	DailyLoc	Category	G-RANK
Marionette-noCond	0.0799	0.0187	0.0089	0.0182	0.0145	0.0429
Marionette-noPartial	0.0771	0.0197	0.0111	0.0202	0.0104	0.0467
Marionette	**0.0664**	**0.0134**	**0.0083**	**0.0144**	**0.0024**	**0.0248**

Table 5: Ablation study in Task-Based Evaluation on MAPE

Method	LocRec	NexLoc	SemLoc	EpiSim
Marionette-noCond	0.2289	0.1274	0.1857	0.0434
Marionette-noPartial	0.1893	0.0878	0.1748	0.0631
Marionette	**0.1862**	**0.0847**	**0.0943**	**0.0352**

5.4 Impact of Different Contexts

In this section, we investigate the impact of different contexts on human mobility using the campus dataset. Our experiments evaluate how varying contexts affect both statistical and distributional similarity metrics, as well as the performance in task-based evaluation. The results are shown in Tables 6 and 7. We observe that different contexts have different extents of impact on human mobility, while Marionette incorporates all contexts, achieves optimal performance across most evaluation scenarios, demonstrating the usefulness of various contexts in modeling human mobility.

5.5 Case Studies on Controllable Generation

To showcase the controllable generation with the partial context using Marionette, we use our trained model on the Campus dataset to generate daily trajectory data in two previously unseen scenarios: the No.8 tropical cyclone warning signal in the morning but no warning in the afternoon (Figure 2), and no warning in the morning but the No.8 tropical cyclone warning signal in the afternoon (Figure 3). We also select the trajectory data of one typical day with no typhoon warning signal as a reference (Figure 4). To better analyze the generated trajectory data, we visualize the three sets of trajectories under transition matrices, shown in heatmaps.

We see that the mobility patterns in the two generated scenarios exhibit significant deviations from the normal condition during active periods of the partial condition of the No.8 tropical cyclone warning signal, with large reductions in crowd movement. In contrast, during periods without warning signals in the two generated scenarios, the mobility patterns remain similar to the normal condition. This shows the manipulability of our Marionette in generating human trajectory data to meet the requirements of user-specified scenarios with fine-grained conditions (partial context).

Table 6: Performance on Statistical and Distributional Similarity Metrics with Different Contexts

Method	Distance	Radius	Interval	DailyLoc	Category	G-RANK
no Almanac	0.0666	0.0140	0.0090	0.0153	0.0045	0.0256
no HourInDay	0.0682	0.0139	0.0085	0.0146	0.0049	0.0256
no Rainstorm	0.0667	**0.0133**	0.0088	0.0158	0.0026	0.0253
no TropicalCyclone	0.0665	0.0134	0.0088	0.0149	0.0053	0.0292
no DayInWeek	0.0677	0.0148	0.0095	0.0201	0.0057	0.0411
no Thunderstorm	**0.0663**	0.0134	0.0088	0.0147	0.0053	0.0292
no TeachingArrangement	0.0689	0.0143	0.0093	0.0168	0.0048	0.0258
all contexts	0.0664	0.0134	**0.0083**	**0.0144**	**0.0024**	**0.0248**

Table 7: Performance in the Task-Based Evaluation on MAPE with Different Contexts

Method	LocRec	NexLoc	SemLoc	EpiSim
no Almanac	0.2052	0.0947	0.1352	0.0553
no HourInDay	0.2047	0.1048	0.1583	0.0460
no Rainstorm	0.1926	0.1018	0.1262	0.0481
no TropicalCyclone	0.2013	0.1064	0.1439	0.0391
no DayInWeek	0.2089	0.1025	0.1534	0.0561
no Thunderstorm	0.1975	0.0990	0.1401	0.0471
no TeachingArrangement	0.2025	0.0964	0.1516	0.0424
all contexts	**0.1862**	**0.0847**	**0.0943**	**0.0352**

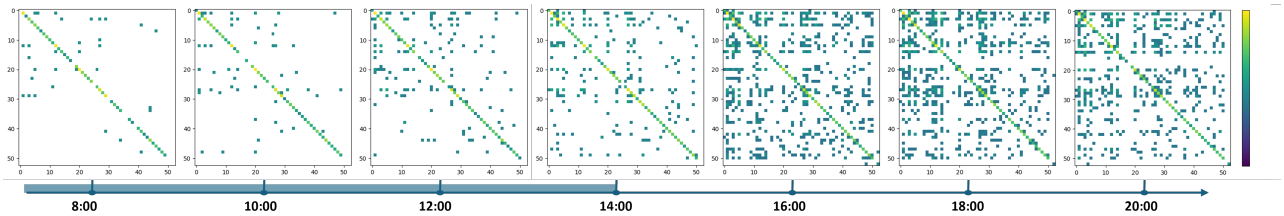


Figure 2: No.8 Tropical Cyclone Warning Signal in the Morning

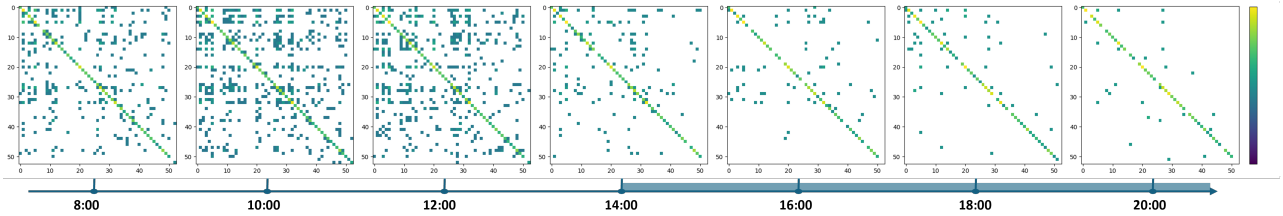


Figure 3: No.8 Tropical Cyclone Warning Signal in the Afternoon

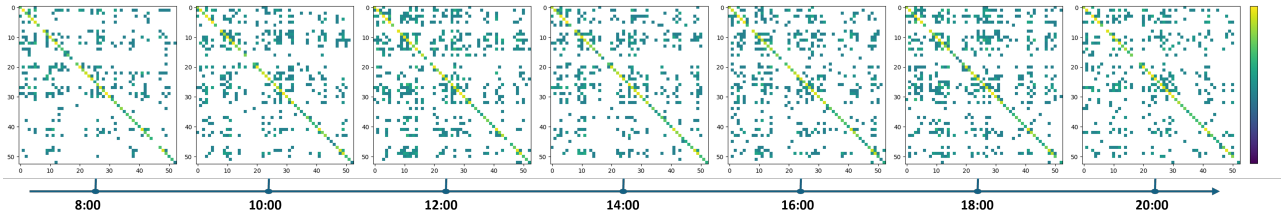


Figure 4: A Typical Day with no Warning Signal

6 Conclusion

In this paper, we study the problem of fine-grained conditional generative modeling of spatiotemporal human trajectory data, and propose Marionette, a **Ma**nipulatable gene**r**ati**i**ve m**o**del for huma**n** traje**c**t**o**ry da**t**a with fin**e**-grained conditions. First, we integrate the mobility-related global and partial contexts using a context alignment method, and then extract both sequence- and event-level conditions. Second, for conditional temporal modeling, we design a conditional TPP following a diffusion-alike TPP scheme to capture the impact of the conditions on both event-triggered shifts and continuous variation of the intensity function of the TPP. Third, for conditional spatial modeling, we design a cascading conditioning mechanism integrated with a discrete diffusion model to capture the cascading dependence of locations on activity categories and further on time. We conduct a thorough evaluation of Marionette on two real-world human trajectory datasets against a sizeable collection of state-of-the-art baselines. Results show that Marionette consistently outperforms the best baselines by 13.96-54.13% on statistical and distributional similarity metrics and by 9.36-40.63% in task-based data utility evaluation. Case studies also demonstrate the superior manipulability of Marionette.

Our future work will extend the context modeling via Large Language Models to further improve its generalizability.

Acknowledgments

This project has received funding from the Science and Technology Development Fund, Macau SAR (0047/2022/A1, 001/2024/SKL), Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC), and UIC Research Grant (UICAG100006). This work was performed in part at SICC which is supported by SKL-IOTSC, University of Macau.

References

[1] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *NIPS* 34 (2021), 17981–17993.

[2] Fred Brauer, Carlos Castillo-Chavez, Zhilan Feng, Fred Brauer, Carlos Castillo-Chavez, and Zhilan Feng. 2019. Models for Influenza. *Mathematical Models in Epidemiology* (2019), 311–350.

[3] Chu Cao and Mo Li. 2021. Generating mobility trajectories with retained Data Utility. In *KDD*. 2610–2620.

[4] David Chapman, Kristina Nilsson, Agneta Larsson, and Agatino Rizzo. 2017. Climatic barriers to soft-mobility in winter: Luleå, Sweden as case study. *Sustainable cities and society* 35 (2017), 574–580.

[5] Chunhua Chen, Yuxin Yang, Hao Yuan, Longbiao Chen, Leye Wang, Bingqing Qu, and Dingqi Yang. 2024. Animating the Crowd Mirage: A WiFi-Positioning-Based Crowd Mobility Digital Twin for Smart Campuses. *IMWUT* 8, 4 (2024), 1–32.

[6] Ricky TQ Chen, Brandon Amos, and Maximilian Nickel. 2020. Neural spatio-temporal point processes. *arXiv:2011.04583* (2020).

[7] Zhenhua Chen, Zhaoya Gong, Shan Yang, Qiwei Ma, and Changcheng Kan. 2020. Impact of extreme weather events on urban human flow: A perspective from location-based service data. *Computers, environment and urban systems* 83 (2020), 101520.

[8] Bangchao Deng, Xin Jing, Tianyue Yang, Bingqing Qu, Dingqi Yang, and Philippe Cudre-Mauroux. 2025. Revisiting Synthetic Human Trajectories: Imitative Generation and Benchmarks Beyond Datasaurus. In *KDD*. 201–212.

[9] Bangchao Deng, Bingqing Qu, Pengyang Wang, Dingqi Yang, Benjamin Fankhauser, and Philippe Cudre-Mauroux. 2025. Replay: Modeling time-varying temporal regularities of human mobility for location prediction over sparse trajectories. *IEEE Transactions on Mobile Computing* (2025).

[10] Bangchao Deng, Dingqi Yang, Bingqing Qu, Benjamin Fankhauser, and Philippe Cudre-Mauroux. 2023. Robust location prediction over sparse spatiotemporal trajectory data: Flashback to the right moment! *ACM Transactions on Intelligent Systems and Technology* 14, 5 (2023), 1–24.

[11] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. 2016. Recurrent marked temporal point processes: Embedding event history to vector. In *KDD*. 1555–1564.

[12] Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. 2020. Learning to simulate human mobility. In *KDD*. 3426–3433.

[13] Bent Fuglede and Flemming Topsoe. 2004. Jensen-Shannon divergence and Hilbert space embedding. In *ISIT*. IEEE, 31.

[14] Song Gao, Jinmeng Rao, Yuhao Kang, Yunlei Liang, and Jake Kruse. 2020. Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSpatial Special* 12, 1 (2020), 16–26.

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

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*. 10696–10706.

[17] Peter E Hart, Nils J Nilsson, and Bertram Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics* 4, 2 (1968), 100–107.

[18] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*. 639–648.

[19] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.

[20] Yayao Hong, Hang Zhu, Tieqi Shou, Zeyu Wang, Liyue Chen, Leye Wang, Cheng Wang, and Longbiao Chen. 2024. STORM: A Spatio-Temporal Context-Aware Model for Predicting Event-Triggered Abnormal Crowd Traffic. *IEEE Transactions on Intelligent Transportation Systems* (2024).

[21] Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. 2019. The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law* 28, 1 (2019), 65–98.

[22] Dou Huang, Xuan Song, Zipei Fan, Renhe Jiang, Ryosuke Shibasaki, Yu Zhang, Haizhong Wang, and Yugo Kato. 2019. A variational autoencoder based generative model of urban human mobility. In *MIPR*. IEEE, 425–430.

[23] Dou Huang, Xuan Song, Zipei Fan, Renhe Jiang, Ryosuke Shibasaki, Yu Zhang, Haizhong Wang, and Yugo Kato. 2019. A variational autoencoder based generative model of urban human mobility. In *MIPR*. IEEE, 425–430.

[24] Renhe Jiang, Xuan Song, Dou Huang, Xiaoya Song, Tianqi Xia, Zekun Cai, Zhaonan Wang, Kyoung-Sook Kim, and Ryosuke Shibasaki. 2019. Deepurbanevent: A system for predicting citywide crowd dynamics at big events. In *KDD*. 2114–2122.

[25] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C González. 2016. The TimeGeo modeling framework for urban mobility without travel surveys. *PNAS* 113, 37 (2016), E5370–E5378.

[26] Wenjun Jiang, Wayne Xin Zhao, Jingyuan Wang, and Jiawei Jiang. 2023. Continuous Trajectory Generation Based on Two-Stage GAN. *arXiv:2301.07103* (2023).

[27] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *ICDM*. IEEE, 197–206.

[28] Vo S Korolyuk, SM Brodi, and AF Turbin. 1975. Semi-Markov processes and their applications. *Journal of Soviet Mathematics* 4, 3 (1975), 244–280.

[29] Shengjie Lai, Nick W Ruktanonchai, Liangcai Zhou, Olivia Prosper, Wei Luo, Jessica R Floyd, Amy Wesolowski, Mauricio Santillana, Chi Zhang, Xiangjun Du, et al. 2020. Effect of non-pharmaceutical interventions to contain COVID-19 in China. *nature* 585, 7825 (2020), 410–413.

[30] Jiahao Li, Huandong Wang, and Xinlei Chen. 2024. Physics-informed neural ode for post-disaster mobility recovery. In *KDD*. 1587–1598.

[31] Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *WWW*. 689–698.

[32] Haowen Lin, John Krumm, Cyrus Shahabi, and Li Xiong. 2024. Controllable Visit Trajectory Generation with Spatiotemporal Constraints. In *2024 IEEE International Conference on Data Mining (ICDM)*. IEEE, 773–778.

[33] Qingyue Long, Huandong Wang, Tong Li, Lisi Huang, Kun Wang, Qiong Wu, Guangyu Li, Yanping Liang, Li Yu, and Yong Li. 2023. Practical synthetic human trajectories generation based on variational point processes. In *KDD*. 4561–4571.

[34] David Lüdke, Marin Biloš, Oleksandr Shchur, Marten Lienen, and Stephan Günnemann. 2023. Add and thin: Diffusion for temporal point processes. *NIPS* 36 (2023), 56784–56801.

[35] Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. 2016. Large-Scale Mobile Traffic Analysis: A Survey. *IEEE Communications Surveys and Tutorials* 18, 1 (2016), 124–161.

[36] Kun Ouyang, Reza Shokri, David S Rosenblum, and Wenzhuo Yang. 2018. A non-parametric generative model for human trajectories.. In *IJCAI*, Vol. 18. 3812–3817.

[37] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv:1205.2618* (2012).

[38] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *WWW*. 811–820.

[39] Chenyang Shao, Fengli Xu, Bingbing Fan, Jingtao Ding, Yuan Yuan, Meng Wang, and Yong Li. 2024. Beyond imitation: Generating human mobility from context-aware reasoning with large language models. *arXiv preprint arXiv:2402.09836* (2024).

[40] Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. 2019. Intensity-free learning of temporal point processes. *arXiv preprint arXiv:1909.12127* (2019).

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

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

[43] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *CIKM*. 1441–1450.

[44] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *WSDM*. 565–573.

[45] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[46] Rajat Verma, Takahiro Yabe, and Satish V Ukkusuri. 2021. Spatiotemporal contact density explains the disparity of COVID-19 spread in urban neighborhoods. *Scientific Reports* 11, 1 (2021), 10952.

[47] Yu Wang, Tongya Zheng, Yuxuan Liang, Shunyu Liu, and Mingli Song. 2024. Cola: Cross-city mobility transformer for human trajectory simulation. In *WWW*. 3509–3520.

[48] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *AAAI*, Vol. 33. 346–353.

[49] Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen Chu. 2017. Modeling the intensity function of point process via recurrent neural networks. In *AAAI*, Vol. 31.

[50] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2017. Deep matrix factorization models for recommender systems.. In *IJCAI*, Vol. 17. Melbourne, Australia, 3203–3209.

[51] Dingqi Yang, Benjamin Fankhauser, Paolo Rosso, and Philippe Cudre-Mauroux. 2020. Location prediction over sparse user mobility traces using rnns. In *IJCAI*. 2184–2190.

[52] Dingqi Yang, Bin Li, and Philippe Cudré-Mauroux. 2016. POIsketch: semantic place labeling over user activity streams. In *IJCAI*. 2697–2703.

[53] Dingqi Yang, Bingqing Qu, Jie Yang, and Philippe Cudre-Mauroux. 2019. Revisiting user mobility and social relationships in lbsns: a hypergraph embedding approach. In *WWW*. 2147–2157.

[54] Dingqi Yang, Bingqing Qu, Jie Yang, and Philippe Cudré-Mauroux. 2020. Lbsn2vec++: Heterogeneous hypergraph embedding for location-based social networks. *IEEE Transactions on Knowledge and Data Engineering* 34, 4 (2020), 1843–1855.

[55] Xiaojie Yang, Hangli Ge, Jiawei Wang, Zipei Fan, Renhe Jiang, Ryosuke Shibasaki, and Noboru Koshizuka. 2024. CausalMob: Causal Human Mobility Prediction with LLMs-derived Human Intentions toward Public Events. *arXiv preprint arXiv:2412.02155* (2024).

[56] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, Vol. 31.

[57] Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, and Yong Li. 2023. Spatio-temporal Diffusion Point Processes. *arXiv:2305.12403* (2023).

[58] Yuan Yuan, Jingtao Ding, Huandong Wang, Depeng Jin, and Yong Li. 2022. Activity trajectory generation via modeling spatiotemporal dynamics. In *KDD*. 4752–4762.

[59] Junyi Zhang, Jiaqi Guo, Shizhao Sun, Jian-Guang Lou, and Dongmei Zhang. 2023. Layoutdiffusion: Improving graphic layout generation by discrete diffusion probabilistic models. In *ICCV*. 7226–7236.

[60] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *CIKM*. ACM, 4653–4664.

[61] Kun Zhou, Yifan Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Diffusion-NAT: Self-Prompting Discrete Diffusion for Non-Autoregressive Text Generation. In *EACL*. 1438–1451.

[62] Yuanshao Zhu, Yongchao Ye, Ying Wu, Xiangyu Zhao, and James Yu. 2023. SynMob: Creating High-Fidelity Synthetic GPS Trajectory Dataset for Urban Mobility Analysis. In *NIPS*.

[63] Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James Yu. 2024. Difftraj: Generating gps trajectory with diffusion probabilistic model. *NIPS* 36 (2024).

[64] Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Qidong Liu, Yongchao Ye, Wei Chen, Zijian Zhang, Xuetao Wei, and Yuxuan Liang. 2024. Controltraj: Controllable trajectory generation with topology-constrained diffusion model. In *KDD*. 4676–4687.

[65] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. 2020. Transformer hawkes process. In *ICML*. PMLR, 11692–11702.

A Baselines

Semi-Markov [28] employs an exponential distribution with Gamma priors for time intervals and utilizes a Dirichlet prior to construct the transition matrix for Bayesian inference. **TimeGeo** [25] designs a statistical EPR model that integrates temporal information, such as weekly home-based tour numbers, dwell rates, and burst rates, to better characterize temporal patterns in human mobility. **RMTPP** [11] leverages Recurrent Neural Networks (RNNs) to jointly capture the dependencies of time and mark (location ID) based on historical information in temporal point processes, enabling more flexible modeling of the joint distribution of time and marks. **ERTPP** [49] employs two distinct RNNs to independently model the timing and marks of events to improve flexibility and performance. **ActSTD** [58] focuses on capturing the continuous spatiotemporal dynamics of human mobility and enhances the dynamic modeling of individual trajectories by utilizing neural ordinary differential equations. **SeqGAN** [56] combines Generative Adversarial Networks (GANs) with reinforcement learning to address the challenge of generating discrete sequences, such as text or event sequences. **MoveSim** [12] is a GAN-based framework that integrates physical regularities and prior knowledge of human mobility in trajectory generation. **VOLUNTEER** [33] incorporates a two-layer VAE model with a TPP to capture the characteristics of human mobility from both group and individual views. **COLA** [47] introduces a model-agnostic transfer framework that separates private and shared modules, adjusting prediction probabilities to reflect city-specific characteristics in the simulation of human trajectories. **DiffTraj** [63] uses a diffusion probabilistic model for continuous location generation where we choose the nearest POI to each generated GPS coordinate as the generated POI. **MIRAGE** [8] integrates an intensity-free neural TPP and a neural EPR model to imitate the human decision-making process in trajectory generation in an autoregressive manner. We further extend MIRAGE with our defined mobility-related contexts by encoding both global and partial contexts into one overall condition, which is then fed as an additional input to further parameterize its autoregressive TPP; this variant is denoted as **Cond-MIRAGE**.

B Task-Based Utility Evaluation Protocol

Following [8], our evaluation protocol includes four typical tasks: location recommendation, next location prediction, semantic location labeling, and epidemic simulation, which model user trajectory data in four aspects, user preferences on locations, sequential mobility patterns, collective traffic patterns, and spatiotemporal contact patterns, respectively. For each task, we choose multiple state-of-the-art techniques to conduct experiments and report the results on multiple metrics, to average out the biases of individual techniques and metrics. We then measure the paired performance discrepancy between the real and generated trajectories using Mean Absolute Percentage Error (**MAPE**) and Mean Squared Percentage Error (**MSPE**), which serve as final benchmarks to assess the ultimate utility of the generated trajectories. The details of the four tasks are as follows: 1) **Location Recommendation Task (LocRec)**: considers five popular recommendation algorithms, i.e., BPR [37], DMF [50], LightGCN [18], MultiVAE [31], and NeuMF [19] and report their performance on Mean Reciprocal Rank@N (MRR@N), Normalized Discounted Cumulative Gain@N (NDCG@N), hit@N (where N = 5 and 10). 2) **Next Location Prediction Task (NexLoc)**: considers five sequence prediction algorithms, i.e., FPMC [38], BERT4Rec [43], Caser [44], SRGNN [48], and SASRec [27], and report their performance on MRR@N, NDCG@N, hit@N (where N = 5 and 10). Note that our LocRec and NexLoc tasks are implemented by RecBole [60]. 3) **Semantic Location Labeling Task (SemLoc)**: assigns a semantic label (i.e., activity category) to a location based on the *collective traffic pattern* of the location, extracted from users' trajectory data [52]. We consider five typical classification algorithms, i.e., Decision Tree, Naive Bayes, K-Nearest Neighbors, Logistic Regression, and Support Vector Machine, and report their performance on Accuracy, F1-Micro, and F1-Macro scores. 4) **Epidemic Simulation Task (EpiSim)**: simulates the epidemic spreading over a contact network characterizing the *spatiotemporal contact patterns* of user trajectories [46]. Following the setting of recent works [12, 29, 58], we adopt the Susceptible–Exposed–Infected–Recovered (SEIR) model. We consider the simulation of COVID-19 using the parameters suggested by [12, 58] and influenza using the parameters suggested by [2], and report the average results of 10 repeated simulations to discount the impact of the random selection of initially exposed individuals. Note that for each individual task, we compare the paired performance discrepancy between the real and generated trajectories using MAPE and MSPE.

Table 8: Performance in Task-Based Evaluation on MSPE

	Campus				Istanbul			
Method	LocRec	NexLoc	SemLoc	EpiSim	LocRec	NexLoc	SemLoc	EpiSim
Semi-Markov	0.8435	0.8461	0.1551	0.5607	0.9663	0.9663	0.1533	0.3937
Time Geo	0.4789	2.5793	0.0976	0.2032	0.7252	0.8981	0.1100	0.6058
RMTPP	0.4526	0.3893	0.1147	0.1164	0.4640	0.3607	0.3024	0.1906
ERTPP	0.3138	0.4529	0.1043	0.0798	0.5643	0.4979	0.1418	0.3155
ActSTD	0.4737	0.4375	0.4594	0.6237	0.6145	0.5527	0.1290	0.4639
SeqGAN	0.3006	0.9174	0.1434	0.6609	2.3803	0.6412	0.1444	0.3379
MoveSim	4.7328	0.2840	0.2433	0.0277	0.7877	0.6819	0.1573	0.0931
VOLUNTEER	0.3971	0.5712	0.3801	0.5257	0.9155	0.8036	0.1718	0.0482
DiffTraj	0.9884	0.2389	0.2032	0.3376	0.7111	0.6854	0.1492	0.2989
COLA	0.5504	0.2658	0.2655	0.5347	0.6129	0.5982	0.1706	0.0909
MIRAGE	0.1257	0.1540	0.0171	0.0141	0.3348	0.3713	0.0447	0.0417
Cond-MIRAGE	0.1469	0.1196	0.0236	0.0065	0.1584	0.2446	**0.0268**	0.0273
Marionette	**0.0366**	**0.0075**	**0.0149**	**0.0021**	**0.1226**	**0.1937**	0.0344	**0.0129**

Table 9: Ablation study in Task-Based Evaluation on MSPE

Method	LocRec	NexLoc	SemLoc	EpiSim
Marionette-noCond	0.0577	0.0173	0.0357	0.0024
Marionette-noPartial	0.0387	0.0079	0.0314	0.0069
Marionette	**0.0366**	**0.0075**	**0.0149**	**0.0021**

C Task-Based Evaluation Performance on MSPE

Table 8 shows performance comparison in the task-based evaluation on MSPE. Similar to the results on MAPE, we observe that Marionette achieves the best performance in most cases. Compared to the best-performing baselines, Marionette achieves 61.35% and 16.97% improvement over the best baselines on the two respective datasets. Table 9 shows the ablation study results in the task-based evaluation on MSPE. Similar to the results on MAPE, we also observe that Marionette outperforms its variants Marionette-noCond and Marionette-noPartial, by 41.31% and 33.09%, respectively.