

Article

Controllable Spatio-Temporal Modeling of Pedestrian Spawn Dynamics for Urban Crowd Geosimulation

Yan Lyu ^{1,*}, Bo Ling ^{1,†} , Weiwei Wu ¹, Xiangxiang Xing ^{1,2} and Peng Wang ³

¹ School of Computer Science and Engineering, Southeast University, Nanjing 211189, China; bo_ling@seu.edu.cn (B.L.); weiweiwu@seu.edu.cn (W.W.); mohatarem@seu.edu.cn (X.X.)

² No. 8511 Research Institute of CASIC, Nanjing 210007, China

³ Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M1 5GD, UK; p.wang@mmu.ac.uk

* Correspondence: lvyany@seu.edu.cn

† These authors contributed equally to this work.

Abstract

Realistic modeling of pedestrian flow in dense public spaces is important for urban crowd geosimulation, mobility analysis, and indoor public-space geo-information modeling. Although prior research has emphasized microscopic agent interactions, higher-level spawn dynamics—governing when and where pedestrians appear—remain less explored, despite their fundamental role in shaping crowd density and flow. Existing approaches often decouple spatial and temporal generation, limiting their ability to capture rich spatio-temporal correlations, and they lack controllability for user-specific scenarios such as high-density environments. In this paper, we propose a Guided Joint Spatio-Temporal Diffusion framework for pedestrian spawn simulation. Our objective is to develop and evaluate a controllable joint spatio-temporal generative model that produces each pedestrian spawn event—its inter-arrival time, origin, and destination—consistent with observed spawn dynamics and a user-specified normalized local spawn-intensity condition. The model addresses the upstream initialization of a crowd simulation, rather than complete trajectory prediction or microscopic interaction simulation, and is evaluated through both next-event accuracy and fixed-horizon controllability. The method leverages spatio-temporal diffusion point processes to jointly model spatial and temporal spawn events, capturing dependencies overlooked by classical and neural point-process-based methods. To support controllable pedestrian-flow generation for geosimulation and downstream applications, we integrate a conditional denoising network with classifier-free guidance, enabling user-specified factors such as crowd density to steer generation. Experiments on the Grand Central dataset demonstrate that our method outperforms strong baselines, reducing temporal error (T-RMSE) by 38% and achieving consistent improvements in spatial and spatio-temporal accuracy. These results show the potential of diffusion-based spatio-temporal modeling for controllable urban crowd geosimulation and pedestrian mobility data generation.

Keywords: pedestrian flow; urban crowd geosimulation; spatio-temporal modeling; geo-information; diffusion point process; controllable generation; crowd simulation



Academic Editors: Wolfgang Kainz and Hartwig H. Hochmair

Received: 16 June 2026

Revised: 1 August 2026

Accepted: 3 August 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Published by MDPI on behalf of the International Society for

Photogrammetry and Remote Sensing.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Modeling pedestrian flow in dense public spaces is an important problem in urban geo-information, mobility analysis, and crowd geosimulation [1,2]. Realistic generation of when and where pedestrians enter a scene supports applications such as indoor spatial analysis,

digital twins, public-space management, and robot navigation [3–5]. For downstream tasks, including the training and evaluation of social navigation algorithms, constructing realistic simulated crowd environments is essential, especially as reinforcement learning (RL) methods are increasingly adopted for enabling robots to interact effectively within such complex settings [6,7].

Classical approaches to crowd simulation, such as the Social Force Model [8] and ORCA [9], rely on handcrafted rules to endow human agents with capabilities like goal-reaching and collision avoidance. To be more realistic, recent advanced data-driven methods are based on imitation learning [3,10–12] or reinforcement learning [13–15], which can replicate the microscopic interaction pattern observed in real-world pedestrian data.

However, the upstream initialization of agents—when they enter a scene and how their origins and destinations are assigned—has received comparatively less attention, despite its direct effect on density and flow patterns. Existing spawn mechanisms are commonly rule-, schedule-, or temporally driven, and therefore do not explicitly learn the coupling between an arrival time and its OD assignment [16–21].

Figure 1 illustrates this upstream spawn process in Grand Central: the time at which a pedestrian appears is coupled with the origin and destination that initiate a later trajectory. This creates two related challenges. First, a decoupled temporal–spatial pipeline can miss correlations between arrivals and the OD structure of a station flow. Second, replaying only the empirical spawning process does not allow a simulation user to construct comparable low- or high-intensity scenarios for analysis, stress testing, or downstream navigation evaluation.

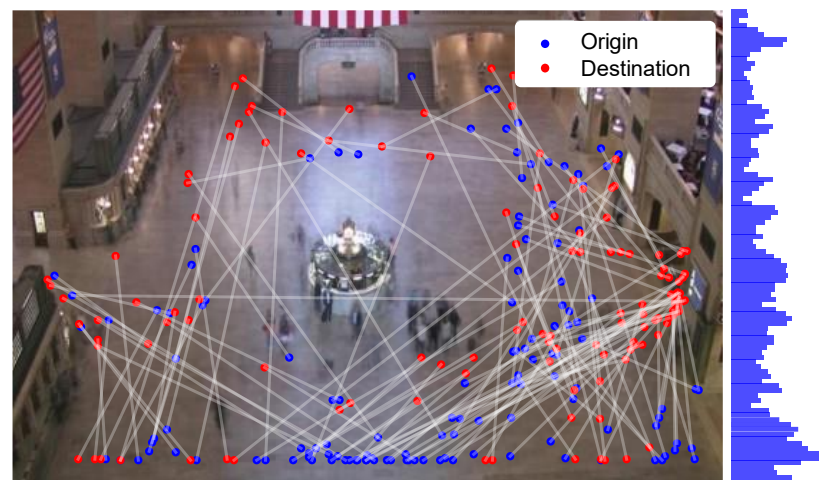


Figure 1. Grand Central spawn dynamics. Blue and red dots denote sampled origins and destinations, respectively; the histogram shows the temporal distribution of spawn events.

In this paper, we introduce a joint spatio-temporal spawning framework based on spatio-temporal diffusion point processes [22] to effectively capture complex spatio-temporal correlations. In addition, we develop a conditional denoising network combined with classifier-free guidance (CFG) [23], allowing a normalized local spawn-intensity condition to guide generation. The study therefore focuses on a clearly delimited task: jointly generating the inter-arrival time, origin, and destination of the next spawn event under a desired local condition. It supplies a realistic initialization process to downstream simulators, rather than replacing trajectory prediction or microscopic interaction models. In summary, our contributions are as follows:

- We propose a framework based on spatio-temporal diffusion point processes that models spatial and temporal pedestrian spawn events jointly, capturing complex spatio-temporal correlations for geo-information modeling.

- We develop a conditional denoising network with classifier-free diffusion guidance, enabling user-specified conditions, such as crowd density, to steer the generation process.
- We demonstrate the effectiveness of our approach through experiments, showing improved realism and controllability in urban crowd geosimulation scenarios.

2. Related Work

Spawn modeling occupies the upstream layer of a crowd-simulation workflow. Rule- and schedule-based mechanisms provide transparent but often manually specified arrivals; classical stochastic processes provide simple temporal variation; and neural temporal point processes improve temporal dependence modeling [16–21]. These approaches differ from microscopic navigation and interaction models, such as social-force, reciprocal-avoidance, imitation-learning, and reinforcement learning methods: the latter determine how agents move after entry, whereas spawn models provide the entry time and OD initialization. The remaining gap is a data-driven mechanism that treats temporal arrival and OD allocation as coupled event attributes while allowing a simulation condition to influence sampling.

From a geo-information perspective, pedestrian spawn modeling is also related to pedestrian movement modeling, spatio-temporal movement analysis, geosimulation of pedestrian flow, and geospatial AI for urban informatics. Prior *IJGI* studies have investigated pedestrian agent-based modeling using scene activity and spatio-temporal trajectory information [24], conceptual frameworks for movement analysis in GIScience [25], and the integration of AI-based geographic simulation models within GISystems [26]. Together, these strands motivate a layered view of crowd geosimulation: schedule- and rule-based mechanisms set arrivals, interaction models evolve agents within the scene, and data-driven spawn generation learns the joint event distribution that links the two. This connection motivates treating spawn events not only as simulation inputs but also as structured spatio-temporal geo-information that describes how crowd flows emerge in indoor public spaces.

Meanwhile, diffusion models have recently emerged as powerful generative frameworks for high-dimensional data, with applications ranging from image synthesis to sequential modeling [27]. Classifier-free guidance (CFG) [23] has further enhanced diffusion models by enabling controllable generation without external classifiers. While CFG has been widely explored in vision and language domains, its potential for spatio-temporal event generation—particularly in crowd simulation—remains largely unexplored. Our work addresses this gap by integrating spatio-temporal diffusion with CFG to achieve realistic and controllable spawn dynamics.

3. Problem Statement

Given a real-world pedestrian dataset, each individual enters the scene at a specific time and navigates from an origin to a destination. We denote the entering time of agent i as $t_i \in \mathbb{R}$, the origin as $o_i \in \mathbb{R}^2$, and the destination as $d_i \in \mathbb{R}^2$. The spatio-temporal spawn event of agent i is defined as $x_i = (t_i, o_i, d_i)$, and the entire dataset's spawn dynamics can be represented as $S = \{x_1, x_2, \dots, x_N\}$, where N is the total number of agents. To enable controllable crowd simulation, we introduce a control variable $c \in \mathbb{R}^n$ representing user-specified factors such as crowd density, where n is the dimensionality of the control variable. In the present implementation, $c = \mathbf{c}_t = [d_{t-2}, d_{t-1}, d_t, d_{t+1}, d_{t+2}] \in \mathbb{R}^5$ is a local temporal profile of normalized frame-level spawn density, rather than five independent semantic attributes. The task of pedestrian spawn simulation is then to generate a sequence of spatio-temporal spawn events that are consistent with the desired control factor c , such that the simulated sequence exhibits realistic crowd dynamics while adhering to the specified

control conditions. During inference, the number of events is not supplied as a fixed input or predicted by a separate count model: events are generated recursively until the next event exceeds the segment's fixed end time, and the retained event count is therefore an emergent outcome. Figure 2 illustrates the overall architecture of the proposed framework.

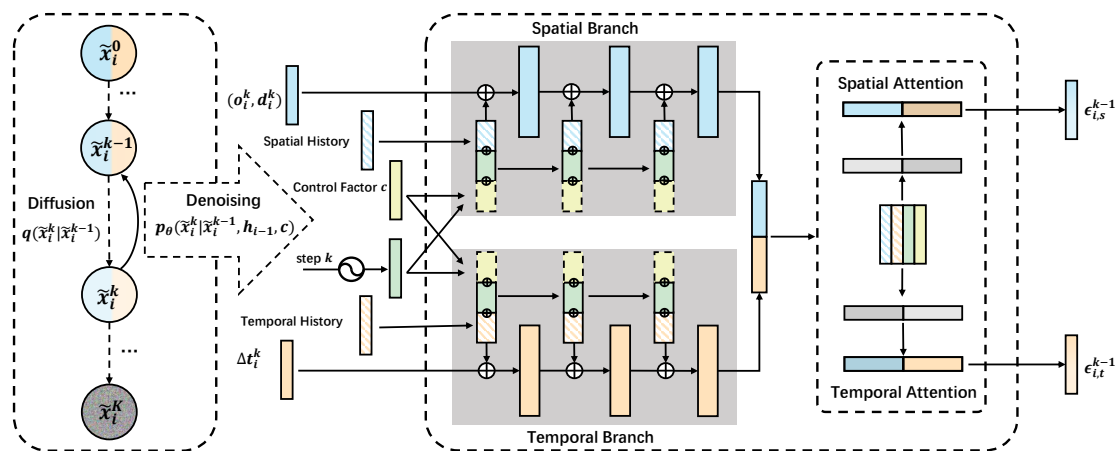


Figure 2. Guided joint spatio-temporal spawn diffusion. The conditional denoiser predicts noise from the previous value, spawn history, and density-control profile. Blue and orange denote spatial and temporal components, respectively; yellow and green denote the control-factor and diffusion-step embeddings.

4. Methodology

We first present a unified framework for modeling and generating pedestrian spawn events as a joint spatio-temporal process. Our method consists of two key components. First, we introduce Joint Spatio-Temporal Spawn Diffusion, which captures the complex dependencies between spatial locations and temporal occurrences of spawn events by leveraging a diffusion-based generative model conditioned on the spawn history. This formulation allows the model to learn the full joint distribution of future events rather than treating spatial and temporal components independently. Second, we enable controllable generation by incorporating classifier-free guidance (CFG) into the denoising network. CFG allows user-specified factors, such as crowd density or other context variables, to effectively guide the generation while maintaining the flexibility of the diffusion process. Together, these components form a coherent framework that can both faithfully model pedestrian-flow dynamics and produce controllable, realistic future spawn sequences.

Operationally, the history encoder summarizes the previously observed spawn events, while the diffusion model generates one noisy candidate next event in reverse. The candidate contains an inter-arrival time and an ordered OD coordinate pair, so the temporal and spatial attributes are exposed jointly to the history-dependent denoising process. The local spawn-intensity condition is embedded alongside these history features and is strengthened at inference through CFG. This design makes the control an input to the next-event distribution rather than a post hoc rule that directly fixes a segment's event count.

4.1. Joint Spatio-Temporal Spawn Diffusion

Instead of decoupling the spatial and temporal components of the generation process as in [21], we model crowd spawning dynamics as a unified spatio-temporal process. Specifically, we employ spatio-temporal diffusion point processes (DSTPPs) [22], which leverage diffusion-based generative modeling to directly capture the joint spatio-temporal distribution of events. We inherit DSTPPs' history-conditioned reverse-diffusion principle for the next event, but adapt the event representation to pedestrian spawning: one event

is an inter-arrival time jointly paired with an OD coordinate vector, $\tilde{x}_i = (\Delta t_i, o_i, d_i)$. The controllable density conditioning and classifier-free guidance are task-specific additions rather than components of the DSTPP backbone.

Spatio-temporal history encoder. To model the conditional probability $p(x|H_t, c)$, $H_t = \{x_j | x_j \in S, t_j < t\}$, we first introduce the spatio-temporal spawn history encoder, which encodes the sequence of spawn history with a Transformer architecture. Given each spawn event $x_i = (t_i, o_i, d_i)$, the temporal part t_i uses sinusoidal positional encoding [28] and the spatial part (o_i, d_i) is projected via a linear embedding. Concretely, the spatial input is the ordered four-dimensional vector (o_x, o_y, d_x, d_y) : coordinate order preserves the origin and destination roles and permits their correlation to be learned jointly, but the current encoder does not use separate role-specific encoders or semantic region constraints. The resulting spatial embedding e^s , temporal embedding e^t , and their combined embedding e^{st} are then processed by the self-attention modules of the Transformer to yield the hidden representation of spatial history $h^s(t)$, temporal history $h^t(t)$ and spatio-temporal history $h^{st}(t)$. Let h_{i-1} denote the collection.

Diffusion and denoising processes. Given the hidden representations of the spawn history together with the control factor c , we employ the diffusion model [27] to capture the joint spatio-temporal distribution of future spawn events. The diffusion model [27] is a deep generative model that learns data distributions by reversing a gradual noising process. In the training, data samples are progressively perturbed into Gaussian noise, and the model is optimized to predict the noise. Once trained, the model can generate realistic samples by iteratively transforming random noise back into structured data.

With a slight transformation, we represent each spawn event as $\tilde{x}_i = (\Delta t_i, o_i, d_i)$, where t_i is replaced by Δt_i , the time interval since the previous spawn event. Specifically, the diffusion process is a Markov process of adding noise step by step, forming a sequence of $(\tilde{x}_i^0, \tilde{x}_i^1, \dots, \tilde{x}_i^K)$, where K is the step number of diffusion, $\tilde{x}_i^0 = \tilde{x}_i$ and $\tilde{x}_i^K$ can be seen as pure Gaussian noise. The diffusion step is formulated as

$$q(\tilde{x}_i^k | \tilde{x}_i^{k-1}) = \mathcal{N}(\tilde{x}_i^k; \sqrt{1 - \beta_k} \tilde{x}_i^{k-1}, \beta_k \mathbf{I}), \quad (1)$$

where $\{\beta_k\}$ is a variance schedule.

The denoising process reconstructs the spawn event $\tilde{x}_i = (\Delta t_i, o_i, d_i)$ from pure Gaussian noise $\tilde{x}_i^K$ to $\tilde{x}_i^0$, with $\tilde{x}_i^0 = \tilde{x}_i$, by iteratively predicting the previous value conditioned on the hidden representations of spawn history and the control factor, which is $p_\theta(\tilde{x}_i^{k-1} | \tilde{x}_i^k, h_{i-1}, c)$. Specifically, we have

$$p_\theta(\tilde{x}_i^{k-1} | \cdot) = p_\theta(\Delta t_i^{k-1} | \cdot) p_\theta(o_i^{k-1}, d_i^{k-1} | \cdot). \quad (2)$$

The entire denoising process to reconstruct the spatio-temporal spawn event is formulated as follows:

$$p_\theta(\tilde{x}_i^{0:K} | h_{i-1}, c) = p_\theta(\tilde{x}_i^K) \prod_{k=1}^K p_\theta(\tilde{x}_i^{k-1} | \tilde{x}_i^k, h_{i-1}, c). \quad (3)$$

4.2. Controllable Generation with Classifier-Free Guidance

To enable controllable generation, we redesign the denoising network in DSTPPs to incorporate the control factor into the spatio-temporal denoising process. Furthermore, we adopt classifier-free guidance [23] during both training and inference, ensuring that the generation is effectively guided by the control factor.

Conditional denoising network. To reconstruct the future spawn event from pure Gaussian noise in a reverse manner, we employ a conditional denoising network

$\epsilon_\theta(\tilde{x}_i^k, h_{i-1}, k, c)$ that predicts the injected noise and progressively restores the spawn event at each diffusion step. In particular, at step k , given the previously denoised value $\tilde{x}_i^k$, the network predicts the temporal noise $\epsilon_{i,t}^{k-1}$ and spatial noise $\epsilon_{i,s}^{k-1}$, conditioned on the current denoising step k , the spawn history h_{i-1} , and the control factor c .

The embedding e_k of the denoising step is obtained by applying a sinusoidal positional encoder [28]:

$$e_k = \text{SinusoidalPosEmb}(k). \quad (4)$$

Similarly, the control factor c is projected through a multilayer perceptron (MLP) to obtain its embedding e_c .

Then the spatial and temporal attention coefficients are computed as follows:

$$\alpha_s = \text{softmax}(W_s[e_k; h_{i-1}; e_c] + b_s), \quad (5)$$

$$\alpha_t = \text{softmax}(W_t[e_k; h_{i-1}; e_c] + b_t), \quad (6)$$

where W_s, W_t, b_s, b_t are learned parameters; the semicolon denotes concatenation. These coefficients adaptively weight spatial and temporal branches.

For the backbone encoding, both the temporal part $\tilde{x}_{i,t}^k = (\Delta t_i^k)$ and spatial part $\tilde{x}_{i,s}^k = (o_i^k, d_i^k)$ are passed through parallel three-layer MLPs separately with shared conditioning:

$$v_s^{(j)} = \sigma(W_s^{(j)}(v_s^{(j-1)} + h_{s,i-1}^{(j)} + e_k^{(j)} + e_c^{(j)})), \quad (7)$$

$$v_t^{(j)} = \sigma(W_t^{(j)}(v_t^{(j-1)} + h_{t,i-1}^{(j)} + e_k^{(j)} + e_c^{(j)})), \quad (8)$$

where $v_s^{(0)} = (o_i^k, d_i^k)$, $v_t^{(0)} = \Delta t_i^k$, $\sigma(\cdot)$ is ReLU and $h_{s,i-1}^{(j)}, h_{t,i-1}^{(j)}, e_k^{(j)}, e_c^{(j)}$ correspond to the linear projections of $h_{s,i-1}, h_{t,i-1}, e_k, e_c$ respectively. Let v_s and v_t denote the final output encoding.

Finally, the network produces spatial and temporal noise predictions by two projection heads:

$$z_s = \sum_{r \in \{s,t\}} \alpha_{s,r} v_r, \quad z_t = \sum_{r \in \{s,t\}} \alpha_{t,r} v_r, \quad \epsilon_{i,s}^{k-1} = g_s(z_s), \quad \epsilon_{i,t}^{k-1} = g_t(z_t), \quad (9)$$

where $\alpha_{s,r}$ and $\alpha_{t,r}$ are the two softmax weights for branch r , and g_s and g_t are separate output heads. Thus, the attention weights aggregate branch representations rather than directly representing predicted noise.

Training and inference with CFG. As noted, the conditional denoising network is conditioned on the denoising step k , spawn history h_{i-1} and control factor c . Among these, our primary focus is the user-specific control factor c , such as crowd density. Simply feeding the control factor into a standard diffusion model may reduce the model's ability to generate the desired outcomes. To address this, we adopt classifier-free guidance (CFG) [23] to strengthen the dependency between the control factor and the generated future spawn events.

Classifier-free guidance is a technique in diffusion models that enables stronger control over the generated output by optionally conditioning on certain factors without using an external classifier. It works by interpolating between conditioned and unconditioned predictions, effectively amplifying the influence of the control signal on the final generated sample. CFG modifies the standard diffusion framework by introducing a simple yet effective change in both the training and inference phases.

Note that the denoising network is defined as $\epsilon_{\theta}(\tilde{x}_i^k, h_{i-1}, k, c)$, where our primary focus is the control factor. During training, we randomly drop the control factor c with probability p_{drop} while keeping all other inputs unchanged:

$$c_{\text{train}} = \begin{cases} c, & \text{with probability } 1 - p_{\text{drop}}, \\ \emptyset, & \text{with probability } p_{\text{drop}}. \end{cases} \quad (10)$$

The loss is computed as standard denoising:

$$\mathcal{L}_{\text{CFG}} = \mathbb{E}_{\tilde{x}_i^0, \epsilon, k} \left[\|\epsilon - \epsilon_{\theta}(\tilde{x}_i^k, h_{i-1}, k, c_{\text{train}})\|^2 \right]. \quad (11)$$

During inference, let $\epsilon_{\theta}(\cdot|c) = \epsilon_{\theta}(\tilde{x}_i^k, h_{i-1}, k, c)$ and $\epsilon_{\theta}(\cdot|\emptyset) = \epsilon_{\theta}(\tilde{x}_i^k, h_{i-1}, k, \emptyset)$ denote the predictions with and without conditioning. The guided prediction is

$$\epsilon_{\text{CFG}}(\cdot, c) = \epsilon_{\theta}(\cdot|\emptyset) + w \cdot (\epsilon_{\theta}(\cdot|c) - \epsilon_{\theta}(\cdot|\emptyset)), \quad (12)$$

where w is the guidance weight controlling the influence of c .

The entries of the control profile are derived from frame-level spawn counts and are min-max normalized to $[0, 1]$ using the same minimum and maximum computed from the training data. We therefore refer to c_t as a normalized local spawn-intensity condition rather than as instantaneous physical occupancy density.

This formulation allows the network to leverage both conditioned and unconditioned predictions, strengthening the effect of the user-specific control factor while retaining the flexibility of the spatio-temporal spawn diffusion process.

5. Experiments

In this section, we evaluate the effectiveness of our proposed approach on modeling and generating pedestrian spawn events. The evaluation has two complementary purposes: the prediction experiment assesses whether the learned next-event distribution preserves observed temporal and OD patterns, while the controllability experiment examines whether a user-specified local spawn-intensity condition produces systematic changes within a fixed simulation horizon. We first compare its spatio-temporal prediction performance against existing baselines, and then investigate the controllability of the generation process under a user-specific control factor.

5.1. Experiment Settings

Dataset. We conduct experiments on the New York Grand Central (GC) dataset [29], which comprises 12,600 pedestrian trajectories, with the densest scene involving nearly 290 individuals. The data are segmented at intervals of 500 frames with a 10% overlap, resulting in variable-length segments. We use 200 segments for training, and 25 segments each for validation and testing.

Baselines. We compare our approach against several baselines. Specifically, nTPP-GMM [21] models the crowd spawn dynamics by employing a neural Temporal Point Process for the temporal component and a Gaussian Mixture Model for the spatial distribution. Poisson-GMM [20] assumes the temporal process follows a standard Poisson process while modeling the spatial distribution with a Gaussian Mixture Model. Empirical-KDE serves as a basic heuristic baseline that relies on random sampling from the ground truth for temporal generation [18,19] and employs Kernel Density Estimation (KDE) for spatial modeling.

Metrics. We evaluate the spatial and temporal performance of different methods using four metrics. Temporal Mean Absolute Error (T-MAE) measures the deviation between predicted and ground-truth time intervals, while Temporal Root Mean Square Error (T-RMSE) provides a variance-sensitive assessment of temporal accuracy. For spatial evaluation, we adopt Spatial Average Euclidean Distance (S-AED), which quantifies the discrepancy between predicted and true spawn origins and destinations. Finally, we introduce Spatio-Temporal Average Euclidean Distance (ST-AED) to jointly assess spatio-temporal prediction quality.

Implementation Details. Our implementation builds upon the open-source code-base of DSTPPs [22]. The conditional model is optimized with AdamW using an initial learning rate of 1×10^{-3} , $(\beta_1, \beta_2) = (0.9, 0.99)$, a five-epoch warm-up, and linear decay to 5×10^{-5} . We use a batch size of 32 and an L_2 noise-prediction objective. The diffusion process uses a cosine noise schedule; the reported conditional-generation experiments use 200 reverse-diffusion sampling steps, classifier-free guidance weight $w = 3.0$, and control-drop probability $p_{\text{drop}} = 0.2$. The Transformer-based history encoder has $d_{\text{model}} = 64$, $d_{\text{mn}} = 256$, $d_{\text{inner}} = 128$, four layers, four attention heads, $d_k = d_v = 16$, and dropout 0.1; the denoising network uses 64 hidden units. The main training and accuracy experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU and an Intel Core i9-10940K CPU @ 3.30GHz with 32 GB of memory; the separate rollout-efficiency benchmark reported below was measured on an NVIDIA GeForce RTX 5060 Laptop GPU.

For density control, d_t denotes the frame-level spawn count normalized with the training-set minimum and maximum, and $\mathbf{c}_t$ follows the local five-frame profile defined in Section 3. It therefore represents local spawn intensity rather than a segment-level peak or spatial occupancy density. In scalar controllability visualizations, every entry of this five-frame profile is assigned the same target level d , yielding a uniform local condition. The sampler clips normalized outputs to the data range, but no post-processing projects OD samples to entrances, exits, walkable regions, or a navigation mesh; spatial plausibility is learned from the Grand Central distribution.

5.2. Experimental Results

We summarize and compare the performance of our method against baseline approaches under the aforementioned spatial and temporal metrics. The comparison isolates the benefit of coupling temporal and OD attributes: Empirical-KDE and Poisson-GMM use simpler independently specified components, whereas nTPP-GMM improves temporal modeling but still combines it with a separate spatial distribution. In addition, we assess the impact of the control factor to examine the effectiveness of guided generation, where the normalized crowd density is used as the control factor for illustration.

5.2.1. Performance Comparison

Table 1 summarizes the results of different methods. As shown, the heuristic baseline Empirical-KDE exhibits high errors, particularly suffering from the highest temporal variance error (T-RMSE) and spatial error (S-AED) due to its simplistic ground-truth sampling and decoupled spatial KDE modeling. Poisson-GMM also performs poorly, yielding the worst T-MAE and ST-AED, as its temporal dynamics are overly simplified by assuming a rigid Poisson process. While incorporating nTPPs substantially improves temporal correlation modeling, nTPP-GMM still underperforms on the spatial metric (S-AED) and the spatio-temporal metric (ST-AED), since it models spatial and temporal dynamics separately and relies on GMMs, which lack the expressiveness to capture fine-grained spatial distributions. Our method obtains the lowest error on all four metrics, including a T-RMSE of 14.62 compared with 23.48 for nTPP-GMM. This pattern suggests that representing the next

spawn event with coupled temporal and OD information is beneficial not only for temporal accuracy but also for the combined spatio-temporal criterion.

Table 1. Performance comparison with baseline methods. The evaluation includes temporal (T-MAE, T-RMSE), spatial (S-AED), and joint spatio-temporal (ST-AED) metrics.

Method	T-MAE	T-RMSE	S-AED	ST-AED
Empirical-KDE	23.50	61.85	42.72	53.51
Poisson-GMM	32.57	50.06	42.32	58.92
nTPP-GMM	9.74	23.48	40.78	44.30
Ours	9.16	14.62	37.81	40.41

Note: Bold values indicate the best performance.

5.2.2. Influence of Controllable Generation

We visualize four cases of controllable generation, each corresponding to the generation of spawn events conditioned on a specific starting event from different segments, and examine the influence of the control factor. Each case has a fixed 500-frame horizon. Starting from the observed initial event, the model samples $(\Delta t, o, d)$ recursively, updates the current time by $t \leftarrow t + \Delta t$, and stops when the next event would exceed the horizon; only events within the horizon are retained. Consequently, the control does not directly set an event count. Higher local density conditions tend to produce shorter sampled inter-arrival times, allowing more events to fit into the same horizon. As shown in Figure 3, the bars use scalar control levels from $d = 0.2$ to $d = 1.0$; for each level, all entries of the local five-frame profile receive the same value rather than being varied separately. From the four cases, we observe that a higher ground-truth spawn count is generally associated with more generated spawn events. Moreover, Figure 4 demonstrates that by tuning the control factor, our method is able to synthesize denser yet still realistic crowds spatially. The spatial panels provide a qualitative counterpart to the count plots: they show that a higher spawn-intensity setting changes the number and distribution of sampled OD pairs, while their plausibility remains empirical rather than guaranteed by hard semantic constraints.

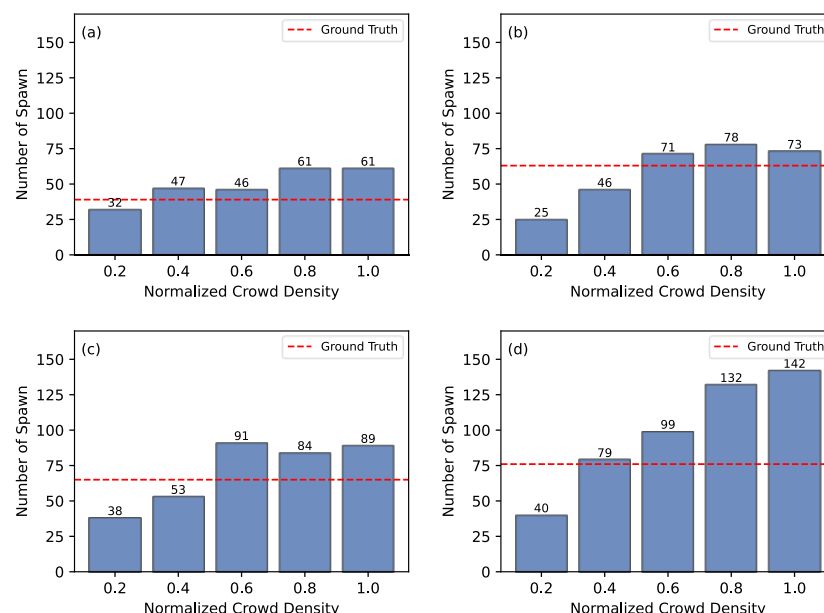


Figure 3. Density-controlled spawn generation in four held-out 500-frame cases. Panels (a–d), initialized from different observed spawn events, have ground-truth counts of 40, 63, 65, and 75, respectively (red dashed lines). Bars show retained generated spawn counts for $d = 0.2$ – 1.0 , applied uniformly across the local five-frame profile.

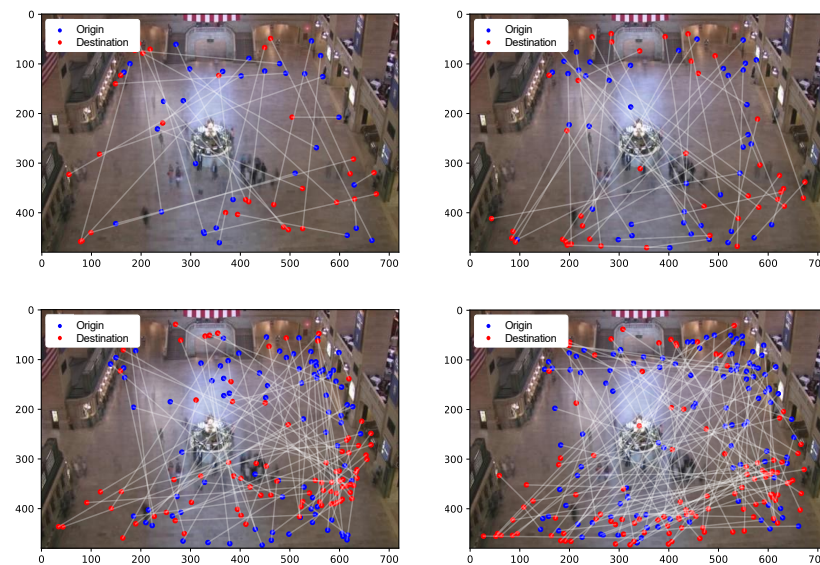


Figure 4. Spatial visualization of density-controlled generation. The **top-left**, **top-right**, **bottom-left**, and **bottom-right** panels show low, moderate-low, moderate-high, and high uniform local spawn-intensity settings, respectively, from the $d = 0.2$ – 1.0 scale. Blue dots, red dots, and gray lines denote sampled origins, destinations, and OD connections.

Inference Efficiency. We benchmarked the saved conditional checkpoint on an NVIDIA GeForce RTX 5060 Laptop GPU using 200 reverse-diffusion steps, $w = 3.0$, and a scalar control level of $d = 0.4$. Across five fixed-horizon 500-frame rollouts after warm-up, inference required 2.32 ± 0.07 s per retained generated spawn and 159.43 ± 104.91 s per full segment; the runs retained 69.4 ± 47.4 spawns on average. Because generation is autoregressive and every next-event sample invokes the reverse-diffusion chain, the current implementation is suited to offline scenario construction and simulation evaluation rather than hard real-time control. Accelerated diffusion samplers are a promising direction for reducing this cost.

6. Conclusions and Discussion

In this work, we presented a diffusion-based framework for controllable pedestrian spawn simulation, which jointly models spatial and temporal dynamics through spatio-temporal diffusion point processes. By integrating a conditional denoising network with classifier-free guidance, our method not only captures complex spatio-temporal dependencies but also enables user-specified controllability, such as adjusting crowd density. Experimental results on the Grand Central dataset demonstrate that our approach outperforms strong baselines in both spatial and temporal accuracy while supporting controllable generation consistent with real-world patterns. These findings highlight the potential of diffusion-based spatio-temporal modeling for controllable urban crowd geosimulation, geo-information modeling, and mobility-analysis applications, while also supporting downstream evaluation tasks such as robot navigation.

To situate the findings relative to existing spawn models, Empirical-KDE provides an empirical but temporally simple construction [18,19], Poisson-GMM imposes a restrictive arrival-process assumption [20], and nTPP-GMM strengthens temporal dependence modeling while retaining a separately fitted spatial mixture [21]. On the Grand Central test set and under the same evaluation protocol, our model reduced T-MAE, T-RMSE, S-AED, and ST-AED by approximately 6%, 38%, 7%, and 9%, respectively, relative to nTPP-GMM, the strongest evaluated baseline. These gains cover temporal accuracy, spatial OD accuracy, and their joint criterion, rather than only one component. They support the hypothesis that exposing

the inter-arrival time and ordered OD vector jointly to the history-conditioned denoiser is beneficial. The consistent improvements across all four metrics demonstrate the effectiveness of the proposed joint spatio-temporal formulation under the present evaluation setting and motivate its further evaluation across diverse datasets and public-space layouts.

Methodologically, the proposed framework adapts the history-conditioned joint next-event generation of DSTPPs [22] to pedestrian OD spawning and adds task-specific spawn-intensity conditioning through classifier-free guidance [23]. Its role is complementary to established microscopic models such as the Social Force Model [8] and ORCA [9]: those models govern movement and interaction after agents enter a scene, whereas our model supplies temporally ordered entry events and OD initialization. The controllability results further show a fixed-horizon effect in which a higher local spawn-intensity condition generally yields shorter sampled inter-arrival times and allows more events before the horizon ends. Thus, the condition steers a conditional event distribution rather than prescribing an exact pedestrian count, making the present model most appropriate for constructing controlled offline simulation scenarios.

Several limitations define the scope of these findings. First, training and evaluation use one station dataset, so transfer to other layouts, demand patterns, and public-space types remains unverified. Second, spatial validity is learned statistically and is not guaranteed by entrance, exit, walkable-region, or navigation-mesh constraints. Map- and region-aware decoding could draw on semantic crowd representations and scene-activity-aware pedestrian modeling [17,24]. Third, the five-dimensional condition is a local temporal profile rather than genuinely multi-attribute control; future extensions could incorporate attributes analogous to configurable crowd profiles [15]. Finally, autoregressive reverse diffusion has a substantial iterative cost. Evaluating accelerated samplers such as denoising diffusion implicit models [30], together with role-aware OD representations and tests across additional public spaces, may improve efficiency, spatial validity, and generalizability.

Author Contributions: Conceptualization, Yan Lyu and Bo Ling; methodology, Yan Lyu and Bo Ling; software, Bo Ling; validation, Yan Lyu, Bo Ling and Peng Wang; formal analysis, Yan Lyu and Bo Ling; investigation, Bo Ling; resources, Weiwei Wu; data curation, Bo Ling; writing—original draft preparation, Yan Lyu and Bo Ling; writing—review and editing, Weiwei Wu, Xiangxiang Xing and Peng Wang; visualization, Bo Ling; supervision, Weiwei Wu and Peng Wang; project administration, Yan Lyu; funding acquisition, Weiwei Wu. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grants 62472093, 62572246, and 62572120 and by the Natural Science Foundation of Jiangsu Province under Grant BK20230024.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The New York Grand Central dataset analyzed in this study is publicly available; see Yi et al. [29] for details.

Acknowledgments: The authors thank all collaborators and reviewers for their constructive feedback.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RL	Reinforcement Learning
TPP	Temporal Point Process

nTPP	Neural Temporal Point Process
GMM	Gaussian Mixture Model
DSTPP	Diffusion Spatio-Temporal Point Process
CFG	Classifier-Free Guidance
MLP	Multilayer Perceptron
GC	Grand Central
T-MAE	Temporal Mean Absolute Error
T-RMSE	Temporal Root Mean Square Error
S-AED	Spatial Average Euclidean Distance
ST-AED	Spatio-Temporal Average Euclidean Distance

References

1. Liu, M.; Li, L.; Li, Q.; Bai, Y.; Hu, C. Pedestrian Flow Prediction in Open Public Places Using Graph Convolutional Network. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 455. <https://doi.org/10.3390/ijgi10070455>.
2. Wang, W.; Osaragi, T. Learning Daily Human Mobility with a Transformer-Based Model. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 35. <https://doi.org/10.3390/ijgi13020035>.
3. Ling, B.; Lyu, Y.; Li, D.; Gao, G.; Shi, Y.; Xu, X.; Wu, W. Socialgail: Faithful crowd simulation for social robot navigation. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 16873–16880.
4. Kruse, T.; Pandey, A.K.; Alami, R.; Kirsch, A. Human-aware robot navigation: A survey. *Robot. Auton. Syst.* **2013**, *61*, 1726–1743.
5. Robla-Gómez, S.; Becerra, V.M.; Llata, J.R.; Gonzalez-Sarabia, E.; Torre-Ferrero, C.; Perez-Oria, J. Working together: A review on safe human-robot collaboration in industrial environments. *IEEE Access* **2017**, *5*, 26754–26773.
6. Chen, C.; Liu, Y.; Kreiss, S.; Alahi, A. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In Proceedings of the 2019 international conference on robotics and automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 6015–6022.
7. Wu, J.; Wang, Y.; Chen, T.; Jiang, J.; Wang, Y.; An, Q.; Yamashita, A. Diversity-aware Crowd Model for Robust Robot Navigation in Human Populated Environment. *IEEE Robot. Autom. Lett.* **2025**, *10*, 6376–6383.
8. Helbing, D.; Molnar, P. Social force model for pedestrian dynamics. *Phys. Rev. E* **1995**, *51*, 4282.
9. Van Den Berg, J.; Guy, S.J.; Lin, M.; Manocha, D. Reciprocal n-body collision avoidance. In Proceedings of the Robotics Research: The 14th International Symposium ISRR, Lucerne, Switzerland, 31 August–3 September 2009; pp. 3–19.
10. Wei, X.; Lu, W.; Zhu, L.; Xing, W. Learning motion rules from real data: Neural network for crowd simulation. *Neurocomputing* **2018**, *310*, 125–134.
11. Zhao, X.; Zhang, J.; Song, W. A Radar-Nearest-Neighbor based data-driven approach for crowd simulation. *Transp. Res. Part C Emerg. Technol.* **2021**, *129*, 103260.
12. Song, X.; Han, D.; Sun, J.; Zhang, Z. A data-driven neural network approach to simulate pedestrian movement. *Phys. A Stat. Mech. Its Appl.* **2018**, *509*, 827–844.
13. Martinez-Gil, F.; Lozano, M.; Fernández, F. Multi-agent reinforcement learning for simulating pedestrian navigation. In Proceedings of the International Workshop on Adaptive and Learning Agents, Taipei, Taiwan, 2 May 2011; pp. 54–69.
14. Lee, J.; Won, J.; Lee, J. Crowd simulation by deep reinforcement learning. In Proceedings of the 11th ACM SIGGRAPH Conference on Motion, Interaction and Games, Limassol, Cyprus, 8–10 November 2018; pp. 1–7.
15. Panayiotou, A.; Kyriakou, T.; Lemonari, M.; Chrysanthou, Y.; Charalambous, P. CCP: Configurable crowd profiles. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–10.
16. Dickinson, P.; Gerling, K.; Hicks, K.; Murray, J.; Shearer, J.; Greenwood, J. Virtual reality crowd simulation: Effects of agent density on user experience and behaviour. *Virtual Real.* **2019**, *23*, 19–32.
17. Kraayenbrink, N.; Kessing, J.; Tutenel, T.; de Haan, G.; Bidarra, R. Semantic crowds. *Entertain. Comput.* **2014**, *5*, 297–312.
18. Liu, B.; Liu, H.; Zhang, H.; Qin, X. A social force evacuation model driven by video data. *Simul. Model. Pract. Theory* **2018**, *84*, 190–203.
19. Bein Fahlander, L.; Mossberg, M. *Simulating People Flow at an Airport; Case Study; Arlanda Airport*: Stockholm, Sweden, 2020.
20. Zhou, B.; Wang, X.; Tang, X. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2871–2878.
21. Kreutz, T.; Mühlhäuser, M.; Guinea, A.S. Whenever, Wherever: Towards Orchestrating Crowd Simulations with Spatio-Temporal Spawn Dynamics. *arXiv* **2025**, arXiv:2503.16639.

22. Yuan, Y.; Ding, J.; Shao, C.; Jin, D.; Li, Y. Spatio-temporal diffusion point processes. In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 6–10 August 2023; pp. 3173–3184.
23. Ho, J.; Salimans, T. Classifier-free diffusion guidance. *arXiv* **2022**, arXiv:2207.12598.
24. Crooks, A.; Croitoru, A.; Lu, X.; Wise, S.; Irvine, J.M.; Stefanidis, A. Walk This Way: Improving Pedestrian Agent-Based Models through Scene Activity Analysis. *ISPRS Int. J. Geo-Inf.* **2015**, *4*, 1627–1656. <https://doi.org/10.3390/ijgi4031627>.
25. Rahimi, S.; Moore, A.B.; Whigham, P.A. Beyond Objects in Space-Time: Towards a Movement Analysis Framework with ‘How’ and ‘Why’ Elements. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 190. <https://doi.org/10.3390/ijgi10030190>.
26. Zhang, P.; Wu, W.; Xue, C.; Shi, S.; Su, F. A New Framework for Integrating DNN-Based Geographic Simulation Models within GISystems. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 361. <https://doi.org/10.3390/ijgi13100361>.
27. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
28. Zuo, S.; Jiang, H.; Li, Z.; Zhao, T.; Zha, H. Transformer hawkes process. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 11692–11702.
29. Yi, S.; Li, H.; Wang, X. Understanding pedestrian behaviors from stationary crowd groups. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3488–3496.
30. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4–8 May 2021.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.