Crowd Dynamics Demand Adaptivity: Self-Adaptive Physics-Informed Neural Network for Crowd Simulation

Ziying Tan
School of Cyber Engineering, Xidian University
Xi'an, China
zytanxd@stu.xidian.edu.cn

Linbo Luo*
School of Cyber Engineering, Xidian University
Xi'an, China
lbluo@xidian.edu.cn

Haiyan Yin
Centre for Frontier AI Research (CFAR), A*STAR
Singapore
yinhaiyan@outlook.com

Yew-Soon Ong
College of Computing and Data Science, Nanyang Technological University
A*STAR
Singapore
asysong@ntu.edu.sg

Wentong Cai
College of Computing and Data Science, Nanyang Technological University
Singapore
aswtcai@ntu.edu.sg

Abstract

Crowd simulation is crucial for urban planning, traffic management, public safety, and immersive environments. A fundamental challenge is capturing adaptive human behaviors that evolve dynamically with social interactions and task demands. Recently, physics-informed neural networks (PINNs) seamlessly integrate interpretable physics-based models with flexible data-driven learning, significantly enhancing simulation realism. However, current PINN-based methods typically rely on rigid representations of pedestrian perceptions and static task priorities of motion planning, limiting their ability to capture real-world social complexities and behavioral adaptability. To this end, we introduce SA-PINN, a novel Self-Adaptive Physics-Informed Neural Network specifically designed for modeling adaptive crowd behaviors. SA-PINN features two innovative adaptive modules: a self-adaptive social perception module, guided by a visual-field physics model to capture context-dependent social interactions dynamically; and a self-adaptive multi-task PINN training module, automatically balancing key motion objectives such as goal-reaching, collision avoidance, and alignment with real data. By jointly enabling perception-level and task-level adaptations within a unified physics-informed framework, SA-PINN generates highly realistic and physically consistent crowd simulations across diverse environmental contexts. Comprehensive evaluations on three real-world datasets (Lane, Cross 90, and GC) reveal that SA-PINN achieves a 29.7% gain in microscopic trajectory accuracy and enhances macroscopic density similarity by 23.5% compared to the best-performing baselines.

*Corresponding author

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754569

CCS Concepts

• **Computing methodologies** → **Modeling and simulation**.

Keywords

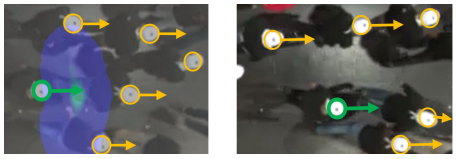
Crowd Simulation, Physics-Informed Neural Networks, Social Perception, Self-Adaptivity

ACM Reference Format:
Ziying Tan, Linbo Luo, Haiyan Yin, Yew-Soon Ong, and Wentong Cai. 2025. Crowd Dynamics Demand Adaptivity: Self-Adaptive Physics-Informed Neural Network for Crowd Simulation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746027.3754569

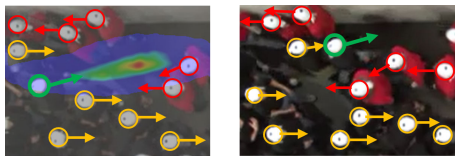
1 Introduction

Crowd simulation is a computational modeling technique that simulates the dynamic behaviors of crowds across various real-world scenarios, finding extensive applications in urban planning, traffic management, and game development. Traditional model-driven crowd simulation approaches [8, 13–15, 23, 24] rely on predefined physical rules to produce specific crowd behavior patterns. However, variations in individual perception and motion-planning capabilities significantly influence crowd behaviors in different contexts. The inherent complexity of heterogeneous crowds presents substantial challenges for accurately modeling crowd dynamics using purely model-driven methods, limiting the realism of simulations. Recent advances in multimedia technology have propelled data-driven methods [4, 16, 21, 32, 34] to the forefront, where crowd movement patterns are learned directly from real-world video data. Such approaches have substantially enhanced simulation realism but remain vulnerable to inaccuracies due to data scarcity and noise. Minor noises in the training data may result in unrealistic simulated behaviors, such as pedestrians overlapping with each other or exhibiting abrupt speed changes.

To overcome these limitations, recent studies have introduced physics-informed neural networks (PINNs) into crowd simulation [5, 17, 19, 29, 31, 36], integrating classical crowd physical models with



(a) Regular attention



(b) Adaptive attention

Figure 1: An illustrative example of regular attention vs. adaptive attention: (a) Regular attention, in which the green-circled individual distributes attention relatively evenly to the surrounding environment and maintains the current direction. (b) Adaptive attention, in which the green-circled individual focuses on available gap space and adjusts direction accordingly.

advanced neural networks to enhance the reliability and realism of simulated behaviors. Existing PINN-based crowd simulations typically replace certain components of physical models with neural networks, enabling end-to-end training by integrating them with other physics modules to learn specific behavioral patterns in crowds. Alternatively, physical rules are incorporated as loss constraints during network training, ensuring that the model generates crowd behaviors consistent with real-world data while adhering to predefined physical principles. These methods often yield more diverse crowd behaviors and mitigate network outputs that could lead to unrealistic crowd dynamics. However, unlike other fields (e.g., fluid dynamics) with well-defined physical laws, crowd systems exhibit complex interactions shaped by individual differences and environmental contexts. For current PINN-based crowd simulation models, physical models of crowd behaviors typically guide the network in a non-adaptive PINN approach. This means that the physical model is parameterized based on empirical assumptions and imposes fixed constraints on the neural network training process. Such a non-adaptive PINN approach overlooks the dynamic evolution of physical priors in different crowd scenarios, which makes it difficult to reproduce subtle variations in crowd motion patterns. Therefore, the key to advancing PINN in crowd simulation is to design self-adaptive physics-informed mechanisms that can capture the adaptive nature of crowd behaviors.

Based on empirical research on crowd behavior [10, 18, 30], the adaptive nature of crowd behavior lies mainly in two key aspects: **social perception** and **motion planning**. First, pedestrians adaptively perceive their social surroundings. As depicted in Fig. 1 (a), an individual usually distributes her or his attention evenly within orderly crowds. In other circumstances (see Fig. 1 (b) for instance), the individual may preferentially focus on a certain area (e.g., navigable gap) of the environment when the social context changes (e.g., encountering opposing crowds). Second, pedestrians make adaptive decisions regarding their motion planning tasks, such as collision avoidance and destination pursuit. For example, pedestrians far from their destination predominantly focus on neighboring states to avoid collisions, while those nearing their destination prioritize goal-directed actions, thus reducing attention to neighbors. Although the adaptive nature of crowd behavior is intuitively clear, explicitly modeling these dynamics remains challenging.

Addressing the aforementioned challenge, we propose a Self-Adaptive Physics-Informed Neural Network (SA-PINN) architecture for crowd simulation that incorporates the learning of adaptivity at both the perception and motion-planning task levels. Inspired by the underlying visual and attention mechanisms in social perception, we propose a **self-adaptive social perception model**, which uses the visual-field physics model to quantify the contexts of social interaction of pedestrians and learns the patterns of attention perception at both the social and individual levels, enabling the adaptivity at the perception level. Moreover, we develop a **self-adaptive multi-task PINN training mechanism**, which integrates a multi-task loss function that jointly optimizes physical (i.e., goal-reaching and collision avoidance) and data consistency tasks. By introducing a task balancing network that adaptively adjusts the loss weights with adversarial training, the model achieves task-level adaptivity. To validate our approach, we performed experiments using three real-world pedestrian datasets, each exhibiting distinct crowd distributions and behaviors. Our model was quantitatively evaluated against recent PINN-based crowd simulation methods at both microscopic and macroscopic scales. The main contributions of our work are summarized as follows:

- We propose a SA-PINN architecture for crowd simulation, which accounts for the dynamic nature of crowd behavior in a physics-informed process.
- We design a self-adaptive social perception model to dynamically model pedestrian-specific social attention patterns.
- We introduce a self-adaptive multi-task PINN training mechanism that dynamically learns task importance, optimizing loss weighting during training.
- Comprehensive experiments show an average 29.7% gain in microscopic trajectory accuracy and 23.5% improvement in macroscopic density similarity compared with the best-performed baselines across three real-world crowd motion datasets, including Lane, Cross 90, and GC.

2 Related Work

As a practical paradigm for integrating domain knowledge with data-driven machine learning, PINNs have gained tremendous momentum in a wide range of domains [12]. Beyond traditional applications in solving partial differential equations (PDEs) [7, 11, 25], PINNs have demonstrated strong potential in multimedia contexts such as motion analysis in videos [22], scene interpretation [6], and human activity recognition [9, 26, 33]. In the field of crowd simulation, PINNs have been applied mainly in two categories: PINNs for **macroscopic** crowd simulation and PINNs for **microscopic** crowd simulation, respectively.

Macroscopic approaches model dense crowds as continuous flows and incorporate fluid dynamics to guide the training of neural networks, enabling the learning of large-scale flow-like motion patterns. Zhou et al. [36] introduced a hydrodynamics-informed network that solves the Navier-Stokes equations via neural networks to simulate dense crowd motion. Pan et al. [19] proposed a

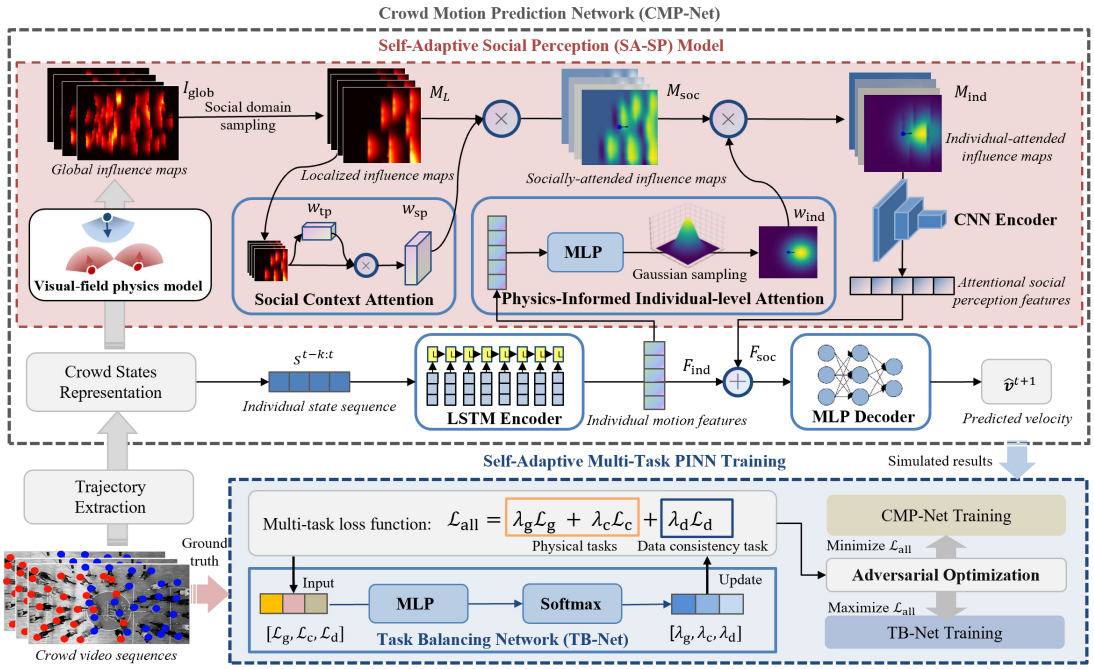


Figure 2: Self-adaptive Physics-Informed Neural Network (SA-PINN) architecture for crowd simulation

reduced-order PINN that decomposes high-order pedestrian flow equations to reduce learning complexity while preserving global dynamics. Although macroscopic PINN excel at modeling collective behaviors, they lack the resolution to capture detail, and individual-level movement, and are often challenging to optimize due to stiff equations, limiting their applicability in generating fine-grained crowd motions.

Microscopic approaches focus on modeling a variety of individual-level behaviors [10, 20]. Zhang et al. [31] leveraged the social force model [8] to simulate physical trajectories, which serve as priors to guide neural network fine-tuning. Yue et al. [29] proposed a neural social physics model that estimates physical parameters through a learnable architecture and integrates them into a motion prediction network. Mo et al. [17] introduced a physics-informed graph neural ODE that encodes pedestrian motion physics via physics-based loss constraints. Chen et al. [5] incorporated the social force model into a diffusion model to guide its denoising process. The microscopic methods can provide a more precise simulation of individual interactions and localized behaviors. However, these methods rely on rigid physical assumptions, such as non-adaptive interaction forces or fixed interpersonal distances, limiting their capabilities to capture the dynamic nature of crowd behaviors.

Our work belongs to the category of designing the PINNs method for microscopic crowd simulation. Different from the existing microscopic approaches, we explicitly incorporate the learning of the adaptivity of crowd behaviors within a PINN architecture. Specifically, we propose a novel self-adaptive PINN framework that unifies perception-level and task-level adaptivity, where we realize attentional social perception in the feature extraction process, and achieve adaptive multi-task balancing in the network training.

3 Methodology

3.1 Overview

We tackle the task of motion prediction for microscopic crowd simulation, where the goal is to learn a predictive model that estimates each pedestrian's future velocity in a dynamic crowd environment. At each time step t, the crowd state is denoted as $S^t = \{s_1^t, s_2^t, \ldots, s_N^t\}$, where N denotes the total number of pedestrians in the scene. The state of the i-th pedestrian is represented as a tuple $s_i^t = (\boldsymbol{p}_i^t, \boldsymbol{v}_i^t, \boldsymbol{d}_i^t)$, where $\boldsymbol{p}_i^t \in \mathbb{R}^2$ is the position, $\boldsymbol{v}_i^t \in \mathbb{R}^2$ is the velocity, and $\boldsymbol{d}_i^t \in \mathbb{R}^2$ is the goal or destination direction. The prediction model f_{pred}, parameterized by θ_{pred}, takes the inputs of the motion history of the i-th agent and its perceived local context to predict the future velocity: $\hat{\boldsymbol{v}}_i^{t+1} = f_{\text{pred}}(s_i^{t-k:t}, C_i^t; \theta_{\text{pred}})$, where $s_i^{t-k:t}$ is the motion history of agent i over the past $k+1$ steps, and C_i^t denotes the perceived context of nearby pedestrians and obstacles. This task poses a profound modeling challenge, requiring simultaneous reasoning over both microscopic interactions (i.e., local responses to nearby agents) and macroscopic dynamics (i.e., global patterns shaped by collective motion). Addressing this complexity requires models that adaptively integrate spatial and temporal information to synchronize individual decisions with the broader dynamics of the crowd.

We propose a self-adaptive physics-informed neural network (SA-PINN) architecture as shown in Fig. 2 that integrates both perception-level and task-level adaptivity guided by effective physics-informed modeling. At its core is the **Crowd Motion Prediction Network (CMP-Net)**, which predicts each pedestrian's future velocity by encoding their motion history and the perceived social context. This context is first captured via a localized influence map, derived using the **visual-field physics model** [18, 27] to represent

each pedestrian's observable region. A social context attention module then adaptively highlights salient interactions by disentangling spatial and temporal cues. To further account for individual variability, a physics-informed attention mechanism generates personalized attention maps using Gaussian distributions parameterized by each agent's velocity and heading, modeling attentional perceptual focus in a physically grounded manner.

For training, SA-PINN employs a multi-task objective that unifies data consistency loss with two physics-informed losses: goal-reaching and collision avoidance, the latter derived from the **optimal reciprocal collision avoidance (ORCA)** model [23], a widely adopted physical model for collision avoidance in crowd simulation. A dedicated **Task Balancing Network (TB-Net)** adaptively reweights these losses through adversarial optimization, allowing the model to self-regulate learning based on the varied importance of motion planning tasks. This design enables SA-PINN to robustly capture the adaptive, physics-consistent behaviors for realistic crowd simulation.

3.2 Self-Adaptive Social Perception Model

We propose a Self-Adaptive Social Perception (SA-SP) model that empowers pedestrian agents to dynamically modulate their social attention in response to varying crowd configurations. Building on the insight that individuals exhibit distinct perceptual behaviors under different social contexts, SA-SP integrates three synergistic components to capture such adaptivity: (1) **Localized Influence Map**, which transforms discrete crowd states into continuous spatial-temporal influence fields using a visual-field physics model, capturing socially coherent interactions within each pedestrian's perceptual range; (2) **Social Context Attention**, which highlights interaction-critical patterns by attending to the most informative regions and time steps in the influence fields; and (3) **Individual-Level Attention**, which predicts pedestrian-specific bivariate Gaussian distributions from motion features to define a physics-informed attention field, where the mean and covariance represent perceptual focus and range, modulating the saliency map to generate individualized attentional perception maps.

3.2.1 Localized Influence Map Generation. To construct the spatial-temporal influence field for each pedestrian, we begin by defining their **social domain** $\mathcal{D}_i$ as a square region of side length l centered at their position. This domain represents the area within which social interactions are considered relevant. Within $\mathcal{D}_i$, the influence of pedestrian i at a spatial location $\boldsymbol{q} \in \mathcal{D}_i$ is computed using a **visual-field physics model** that accounts for both distance decay and directional alignment:

$$I_i(\boldsymbol{q}, \boldsymbol{p}_i, \boldsymbol{v}_i) = exp(-\frac{\|\boldsymbol{q} - \boldsymbol{p}_i\|}{\delta}) \cdot \varphi(\boldsymbol{q} - \boldsymbol{p}_i, \boldsymbol{v}_i),\ q \in \mathcal{D}_i, \quad (1)$$

where $\boldsymbol{q}$ denotes a spatial location within the domain, δ is the distance sensitivity parameter controlling the decay rate of influence over distance, and $\varphi(\cdot)$ is the cosine similarity between the vector from the pedestrian to $\boldsymbol{q}$ relative to its velocity direction:

$$\varphi(\alpha, \beta) = \left(\frac{\alpha \cdot \beta}{\|\alpha\|\|\beta\|}\right). \quad (2)$$

By aggregating the influence of all N pedestrians in the scene, the global influence field $I_{\text{glob}}(\boldsymbol{q})$ and the individual social domain feature $\hat{I}_i(\boldsymbol{q})$ are defined as:

$$I_{\text{glob}}(\boldsymbol{q}) = \sum_{i \in N} I_i(\boldsymbol{q}, \boldsymbol{p}_i, \boldsymbol{v}_i); \quad \hat{I}_i(\boldsymbol{q}) = I_{\text{glob}}(\boldsymbol{q}) - I_i(\boldsymbol{q}, \boldsymbol{p}_i, \boldsymbol{v}_i). \quad (3)$$

The social domain feature $\hat{I}_i(\boldsymbol{q})$ encodes rich, physics-informed spatial cues about nearby agents' motion and proximity, offering a personalized spatial representation of the dynamic social context.

We further extend the social domain feature into the temporal dimension by modeling how pedestrians anticipate future interactions based on short-term motion physics. Assuming a short-term constant velocity, the **expected position** after Δt can be estimated as $\boldsymbol{p}_i^{t+\Delta t} = \boldsymbol{p}_i^t + \boldsymbol{v}_i^t \cdot \Delta t$, from which we derive the corresponding **expected social domain** $\mathcal{D}_i^{t+\Delta t}$ and its associated feature map $\hat{I}_i^{t+\Delta t}(\boldsymbol{q})$ using the same visual-field formulation. By sampling over τ future steps and discretizing each social domain into a spatial grid of size $h \times w$, we construct the **spatial-temporal localized influence maps** $M_L = \{m_L^t, m_L^{t+\Delta t}, \ldots, m_L^{t+(\tau-1)\Delta t}\} \in \mathbb{R}^{\tau \times h \times w}$. This spatial-temporal representation mitigates artifacts arising from discrete feature scale differences and provides a smoother, more informative input to downstream learning modules.

3.2.2 Social Context Attention. We propose a social context attention mechanism that captures the dynamic and diverse influence of surrounding agents by adaptively highlighting socially salient patterns in the localized influence map. Specifically, it comprises two complementary components: temporal attention and spatial attention, operating over the influence maps $M_L \in \mathbb{R}^{\tau \times h \times w}$, which encodes crowd interaction signals across time and space.

The **temporal attention** module focuses on identifying time frames that carry crucial behavioral cues. It performs channel-wise attention along the temporal dimension to enhance salient moments in the influence map:

$$w_{\text{tp}} = f_\sigma\Big(MLP(AvgPool(M_L) + MLP(MaxPool(M_L))\Big), \quad (4)$$

where we first extract global temporal statistics by applying spatial $AvgPool(\cdot)$ and $MaxPool(\cdot)$ operations independently at each time step, resulting in descriptors of shape $\tau \times 1 \times 1$. These descriptors are then processed through separate MLP layers, summed, and passed through a sigmoid activation function $f_\sigma(\cdot)$ to yield temporal attention weights $w_{\text{tp}} \in \mathbb{R}^{\tau \times 1 \times 1}$. Subsequently, the refined temporal features are obtained by applying these weights to the original map as follows: $\widehat{M}_L = M_L \otimes w_{\text{tp}}$, where $\otimes$ denotes the element-wise multiplication with broadcasting.

The **spatial attention** further highlights socially salient regions within each frame that are likely to influence a pedestrian's trajectory. Using the temporally-enhanced maps $\widehat{M}^L$, we derive the spatial attention $w_{\text{sp}} \in \mathbb{R}^{1 \times h \times w}$ as follows:

$$w_{\text{sp}} = f_\sigma\Big(Conv2d\big(\big[AvgPool(\widehat{M}^L), MaxPool(\widehat{M}^L)\big]\big)\Big), \quad (5)$$

where $AvgPool(\cdot)$ and $MaxPool(\cdot)$ denote pooling operations along the temporal dimension, each producing spatial maps of shape $1 \times h \times w$. These pooled spatial maps are concatenated along the channel dimension and passed through a convolutional layer followed by sigmoid activation. Finally, we apply the spatial attention across

all time steps, producing the **socially-attended influence maps** $M_{\text{soc}} = \widehat{M}_L \otimes w_{\text{sp}}$, where $M_{\text{soc}} \in \mathbb{R}^{\tau \times h \times w}$.

By sequentially applying temporal and spatial attention to the localized influence map, our social context attention module allows the model to adaptively attend to when and where social interactions are most influential, offering a rich, socially-aware representation that supports high-fidelity understanding in complex crowd simulation environments.

3.2.3 Physics-Informed Individual-Level Attention. Effective crowd modeling requires agents to adaptively perceive social context through the lens of their own goals and dynamics, enabling personalized responses to surrounding interactions. To this end, we propose a physics-informed individual-level attention mechanism that adaptively interprets social context based on each agent's motion state, producing personalized attention aligned with physical and behavioral plausibility.

Specifically, we model each agent's perceptual focus as a Gaussian distribution over the social domain, parameterized by its own motion features. This physics-informed formulation captures uncertainty and directional sensitivity, allowing attention to adaptively align with the agent's velocity and intent. Formally, given individual motion features F_{ind} extracted via an LSTM encoder from the pedestrian motion sequence, we learn a dedicated Gaussian distribution parameterized by the mean $\boldsymbol{\mu}$ and a covariance matrix Σ that governs the spatial spread of attention. To guarantee that Σ remains positive definite, we constructed it using a lower-triangular matrix A_L with learnable elements $a_{1,1}, a_{2,1}, a_{2,2}$, following the Cholesky decomposition theory [3]:

$$A_L = \begin{bmatrix} exp(a_{1,1}) & 0 \\ a_{2,1} & exp(a_{2,2}) \end{bmatrix}, \quad \Sigma = A_L A_L^T. \tag{6}$$

The resulting Gaussian distribution over spatial coordinates $\boldsymbol{q} \in \mathcal{D}$ is defined as:

$$\mathcal{N}(\boldsymbol{q}; \boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi\sqrt{\Sigma}} exp\Big(-\frac{1}{2}(\boldsymbol{q}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{q}-\boldsymbol{\mu}) \Big). \tag{7}$$

We evaluate this distribution on a uniform spatial grid of resolution $h \times w$, yielding an individual-level attention map $w_{\text{ind}} \in \mathbb{R}^{1 \times h \times w}$. This attention map is further applied to the social-level saliency maps M_{soc} (from Sec 3.2.2) to produce the final **individual-attended influence maps** $M_{\text{ind}} \in \mathbb{R}^{\tau \times h \times w}$:

$$M_{\text{ind}} = M_{\text{soc}} \otimes w_{\text{ind}}. \tag{8}$$

This adaptive physics-informed attention mechanism allows each agent to selectively focus on socially relevant regions based on its own motion state, facilitating personalized and context-aware modeling. The resulting individual-attended influence maps M_{ind} are encoded via a CNN to generate compact attentional social perception features F_{soc}, which are then concatenated with individual motion features F_{ind} and processed through a decoder to predict future velocities:

$$\hat{\boldsymbol{v}}^{t+1} = f_{\text{dec}}([F_{\text{ind}}, F_{\text{soc}}]). \tag{9}$$

3.3 Self-Adaptive Multi-Task PINN Training

Existing crowd simulation methods predominantly focus on motion planning by directly optimizing predicted velocities (Eq. 9), often neglecting the underlying physical constraints essential for generating coherent and socially plausible crowd behaviors. Building on this insight, we reformulate crowd simulation as **a multi-task learning problem**. Here, the **tasks** refer to both the **physical tasks** that real pedestrians consider in motion planning, and the **data consistency task** that ensures alignment between simulated results and real data. We consider two main physical tasks of pedestrians: (1) goal reaching and (2) collision avoidance, which promotes intention-aware and socially compliant motion. To reflect the adaptive nature of real-world pedestrian behavior in social contexts, we propose an adversarial multi-task training framework, where a Task Balancing Network (TB-Net) dynamically adjusts task weights based on contextual difficulty and inter-task conflict, enabling the model to self-regulate learning priorities throughout training.

3.3.1 Multi-task Training Loss. Given the predicted crowd states $\hat{S}^{\mathcal{T}}$ and the ground truth crowd states $S^{\mathcal{T}}$ over the simulation period $\mathcal{T} = (1, 2, .., T)$, the multi-task loss function is formulated as:

$$\mathcal{L}_{\text{all}} = \sum_{t \in \mathcal{T}} \Big[\lambda_{\text{g}} \mathcal{L}_{\text{g}}(\hat{S}^t) + \lambda_{\text{c}} \mathcal{L}_{\text{c}}(\hat{S}^t) + \lambda_{\text{d}} \mathcal{L}_{\text{d}}(\hat{S}^t, S^t) \Big], \tag{10}$$

where $\lambda_{\text{g}} = (\lambda_{\text{g}}^1, ..., \lambda_{\text{g}}^T), \lambda_{\text{c}} = (\lambda_{\text{c}}^1, ..., \lambda_{\text{c}}^T), \lambda_{\text{d}} = (\lambda_{\text{d}}^1, ..., \lambda_{\text{d}}^T)$ are time-dependent task weights that are dynamically adjusted by a learnable loss-balancing network during training.

Given that $s_i^t = (\boldsymbol{p}_i^t, \boldsymbol{v}_i^t, \boldsymbol{d}_i^t)$ is the ground truth state of i-th pedestrian at time t, where $\boldsymbol{p}_i^t$ represents the **position** vector, $\boldsymbol{v}_i^t$ represents the **velocity** vector, and $\boldsymbol{d}_i^t$ is the **displacement** vector from the predicted position to the destination. Correspondingly, $\hat{s}_i^t = (\hat{\boldsymbol{p}}_i^t, \hat{\boldsymbol{v}}_i^t, \hat{\boldsymbol{d}}_i^t)$ denotes the predicted state of pedestrian i at time t. The goal loss $\mathcal{L}_{\text{g}}(\hat{S}^{\mathcal{T}}) \in \mathbb{R}^T$ corresponds to the **goal-reaching task**, and is defined as:

$$\mathcal{L}_{\text{g}}(\hat{S}^{\mathcal{T}}) = \left\{ \sum_{i \in N} exp(-|\hat{\boldsymbol{d}}_i^t|^2) \cdot \varphi(\hat{\boldsymbol{d}}_i^t, \hat{\boldsymbol{v}}_i^t) \right\}_{t=1}^{T}, \tag{11}$$

where $\varphi(\cdot)$ measures the cosine similarity between the $\hat{\boldsymbol{d}}_i^t$ and $\hat{\boldsymbol{v}}_i^t$. The exponential decay term $exp(-\|\hat{\boldsymbol{d}}_i^t\|^2)$ imposes stricter constraints on pedestrians closer to their destination, while progressively relaxing these constraints for those further away.

The collision loss $\mathcal{L}_{\text{c}}(\hat{S}^{\mathcal{T}}) \in \mathbb{R}^T$ for **collision avoidance task** is formulated as:

$$\mathcal{L}_{\text{c}}(\hat{S}^{\mathcal{T}}) = \left\{ \sum_{i \in N} \sum_{i \neq j} max(0, VO_{i|j}^{\tau} \cdot (\hat{\boldsymbol{v}}_i^t - \hat{\boldsymbol{v}}_i^{t-1})) \right\}_{t=1}^{T}, \tag{12}$$

where $VO_{i|j}^{\tau}$ denotes the velocity obstacle (VO) region of pedestrian i induced by pedestrian j over a time window τ, as defined in the ORCA model [23]. Specifically, the $VO_{i|j}^{\tau}$ for pedestrian i induced by pedestrian j over a time window τ is defined as:

$$VO_{i|j}^{\tau} = \left\{ \boldsymbol{v} \,\middle|\, \exists t \in [0, \tau] \text{ such that } t\boldsymbol{v} \in \mathcal{B}(\boldsymbol{p}_j - \boldsymbol{p}_i, r_i + r_j) \right\},$$

where $\mathcal{B}(\boldsymbol{p}, r) = \left\{ \boldsymbol{q} \in \mathbb{R}^2 \,\middle|\, \|\boldsymbol{q} - \boldsymbol{p}\| < r \right\}$ denotes an open ball (disc) centered at $\boldsymbol{p}$ with radius r. The operation $VO_{i|j}^{\tau} \cdot (\hat{\boldsymbol{v}}_i^t - \hat{\boldsymbol{v}}_i^{t-1})$ conceptually evaluates if the velocity change vector points into the VO region. The $max(0, \cdot)$ function penalizes only those velocity

changes that intrude into the VO region, while safe or neutral changes incur no penalty. This loss thus quantifies violations of collision avoidance constraints, serving as a proxy for measuring collision risk.

Finally, the data loss $\mathcal{L}_{\mathrm{d}}(\hat{S}^{\mathcal{T}}, S^{\mathcal{T}}) \in \mathbb{R}^T$ for **data consistency task** is computed as:

$$\mathcal{L}_{\mathrm{d}}(\hat{S}^{\mathcal{T}}, S^{\mathcal{T}}) = \left\{ \sum_{i \in N} \|\hat{\boldsymbol{p}}_i^t - \boldsymbol{p}_i^t\|_2 + \|\hat{\boldsymbol{v}}_i^t - \boldsymbol{v}_i^t\|_2 \right\}_{t=1}^{T}, \tag{13}$$

which measures discrepancies between predicted and ground truth states through deviations in pedestrian positions and velocities.

3.3.2 Adaptive Task Balancing with Adversarial Training. We introduce a **Task Balancing Network** (TB-Net) parameterized by θ_{task}, to adaptively learn the task and time-dependent weights $\boldsymbol{\lambda}$ for optimizing the multi-task losses ($\mathcal{L}_{\mathrm{g}}, \mathcal{L}_{\mathrm{c}}, \mathcal{L}_{\mathrm{d}}$):

$$\boldsymbol{\lambda} = f_{\mathrm{task}}\big(\mathcal{L}_{\mathrm{g}}, \mathcal{L}_{\mathrm{c}}, \mathcal{L}_{\mathrm{d}}; \theta_{\mathrm{task}}\big), \quad \text{with} \quad \boldsymbol{\lambda} = [\lambda_{\mathrm{g}}, \lambda_{\mathrm{c}}, \lambda_{\mathrm{d}}] \in \mathbb{R}^{3T}. \tag{14}$$

During training, the TB-Net and the CMP-Net are jointly optimized via adversarial training:

$$\min_{\theta_{\mathrm{pred}}} \max_{\theta_{\mathrm{task}}} \Big(\lambda_{\mathrm{g}} \cdot \mathcal{L}_{\mathrm{g}} + \lambda_{\mathrm{c}} \cdot \mathcal{L}_{\mathrm{c}} + \lambda_{\mathrm{d}} \cdot \mathcal{L}_{\mathrm{d}}\Big), \tag{15}$$

where θ_{pred} denotes the parameters of CMP-Net. CMP-Net is updated via gradient descent, while TB-Net follows gradient ascent to modulate loss weighting. Given the initial state S^0, and initialized weights $(\lambda_{\mathrm{g}}, \lambda_{\mathrm{c}}, \lambda_{\mathrm{d}})^0$, CMP-Net predicts future crowd state $\hat{S}^{\mathcal{T}}$ over horizon $\mathcal{T}$ in an autoregressive manner. The multi-task losses are computed per Equations 11, 12 and 13, and combined into $\mathcal{L}_{\mathrm{all}}$ using current weights.

TB-Net then updates the loss weights based on the latest multi-task losses ($\mathcal{L}_{\mathrm{g}}, \mathcal{L}_{\mathrm{c}}, \mathcal{L}_{\mathrm{d}}$) to reflect the relative difficulty of each task. These updated weights are applied in the next training epoch to recompute $\mathcal{L}_{\mathrm{all}}$.To stabilize training, we adopt an alternating strategy and update CMP-Net more frequently than TB-Net. This adversarial training framework allows TB-Net to guide CMP-Net by emphasizing underperforming tasks, thereby facilitating a balance of crowd motion across goal-reaching, collision avoidance, and data consistency tasks. The detailed training algorithm is in part (a) of the supplementary material [1].

4 Experiments

4.1 Experimental Setup

Crowd Datasets. To evaluate the method's ability to simulate complex crowd behaviors, we conducted crowd simulation experiments on three public datasets: the Lane dataset [1], the Cross 90 dataset [2], and the Grand Central Station (GC) dataset [35]. These datasets vary significantly in their spatial scales, pedestrian densities, and behavioral patterns, enabling a comprehensive assessment of the generalization ability of our model. Specifically, the Lane dataset captures bidirectional pedestrian counter-flow scenarios with densities surpassing $1.5m^{-2}$. The Cross 90 dataset illustrates complex bidirectional cross-flow dynamics at a crossing angle of 90°, exhibiting maximum densities exceeding $3m^{-2}$. In contrast, the GC dataset, recorded in a real-world transportation hub, represents more freely formed pedestrian movements, with an average density below $0.5m^{-2}$. Detailed descriptions of these datasets can be found in part (b) of the supplementary material.

Baselines. We selected three latest PINN-based methods for crowd simulation as baselines: NSP [29], PCS [31], and SPDiff [5]. NSP embeds an explicit crowd behavior physical model with learnable parameters into deep neural networks. PCS introduces an interactive mechanism, where synthetic training data generated from physical models is used to train neural networks, subsequently applying symbolic regression on trained models to refine the physical representations. SPDiff integrates the classical social force model within a diffusion-based neural network, embedding a physics-inspired crowd interaction module to steer the model's denoising trajectory. Additional implementation details regarding these baseline methods can be found in part (c) of the supplementary material.

Evaluation Metrics. To assess model performance, we employed evaluation metrics at both microscopic and macroscopic levels. At the microscopic level, we quantified trajectory prediction accuracy and realism using Mean Absolute Error (MAE), Optimal Transport (OT), and Maximum Mean Discrepancy (MMD) metrics to evaluate the similarity between simulated and ground-truth pedestrian trajectories. Additionally, we defined and computed the difference in collision counts (Col-) between simulated scenarios and ground truth data to measure the authenticity of pedestrian interactions. At the macroscopic level, we adopted the velocity vorticity difference (VD) [28] and density distribution similarity (DDS) at the macroscopic level to assess the consistency between the simulated crowd movement patterns and the ground truth. The specific explanation and calculation details of the metrics are in part (d) of the supplementary material.

Experiment Settings. All experiments were conducted on a computing server equipped with an Intel i9-12900K CPU and an NVIDIA GeForce RTX 3090 GPU. The simulation horizon was consistently set to $T = 10$ time steps during training. The CMP-Net was pretrained with initial multi-task loss weight $(\lambda_{\mathrm{g}}, \lambda_{\mathrm{c}}, \lambda_{\mathrm{d}})^0 = (0.3, 0.3, 0.4)$. The Adam optimizer was utilized with learning rates of $1e^{-3}$, $3e^{-3}$, and $5e^{-3}$ for the Lane, Cross 90, and GC datasets, respectively. These hyperparameters were determined through preliminary experimental validation across multiple training trials. Additional details regarding parameter configurations and training settings are in part (e) of the supplementary material.

4.2 Experiment Results

As shown in Table 1, we present a comparative evaluation of SA-PINN against baseline methods across three real-world crowd datasets. SA-PINN achieved an average improvement of 29.7% in microscopic trajectory accuracy (MSE) over the best-performing baselines across the three datasets. Specifically, on the Lane dataset, SA-PINN achieves relative improvements ranging from 14.2% to 33.5% on microscopic evaluation metrics (MSE, MAE, OT, and MMD) compared to the best-performing baseline. Similarly, on the Cross 90 dataset, these microscopic metrics exhibit improvements between 12.9% and 25.6%, while on the GC dataset, the microscopic metrics improve by 8.9% to 42.6%. Notably, our approach substantially reduces collision occurrences by 26.6% to 86.7% across all

[1] https://github.com/ziyingTan/Crowd-SA-PINN

Table 1: Comparative performance of simulation models at three crowd datasets

Datasets	Methods	Micro-metrics					Macro-metrics	
		MSE (↓)	MAE (↓)	OT (↓)	MMD (↓)	Col- (↓)	VD (↓)	DDS (↑)
Lane	NSP [29]	0.783	0.503	0.417	0.060	334	0.174	0.461
	PCS [31]	0.963	0.582	0.521	0.065	362	0.357	0.418
	SPDiff [5]	0.248	0.339	0.226	0.014	148	0.201	0.449
	SA-PINN (Ours)	**0.165**	**0.291**	**0.152**	**0.011**	**62**	**0.089**	**0.474**
Cross 90	NSP [29]	0.465	0.465	0.271	0.023	172	0.130	0.282
	PCS [31]	0.888	0.637	0.573	0.043	588	0.187	0.460
	SPDiff [5]	0.587	0.533	0.381	0.030	332	0.289	0.302
	SA-PINN (Ours)	**0.405**	**0.403**	**0.204**	**0.018**	**128**	**0.058**	**0.494**
GC	NSP [29]	2.454	0.966	1.506	0.013	572	0.250	0.317
	PCS [31]	2.593	1.036	1.546	0.012	466	0.272	0.295
	SPDiff [5]	2.747	1.052	1.606	0.010	830	0.364	0.319
	SA-PINN (Ours)	**1.409**	**0.880**	**1.262**	**0.008**	**110**	**0.211**	**0.511**

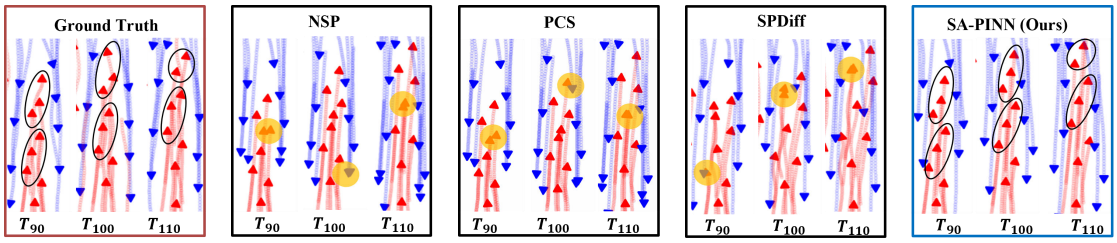


Figure 3: Snapshots of microscopic crowd behaviors of simulated results and ground truth data.

datasets, exhibiting its effectiveness in mitigating unrealistic crowd behaviors through adaptive physics-informed constraints. At the macro-level, SA-PINN yielded average improvements of 39.9% in vorticity and 29.7% in density similarity compared to the strongest baselines. Meanwhile, our method shows greater improvements in the Cross 90, where crowd density is higher and crowd behaviors are more complex. The compared baseline methods have limited generalization capabilities with static physical priors, which struggle under high-density conditions. In contrast, our approach leverages adaptive mechanisms to dynamically adjust motion patterns in response to varying interaction contexts, enabling more robust modeling of complex crowd behaviors.

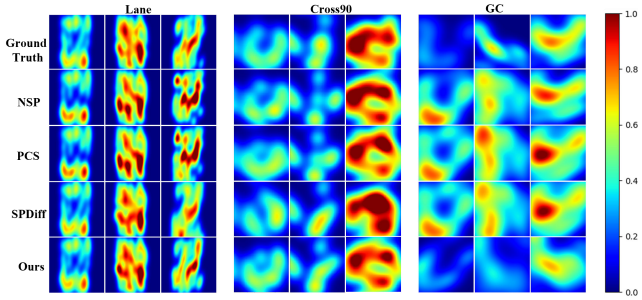


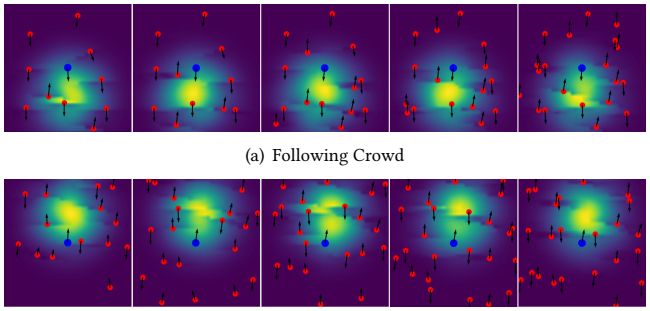
Figure 4: Density distribution maps of simulated results and ground truth data.

To intuitively illustrate the performance differences among methods, we visualize the crowd simulation results from both macroscopic and microscopic perspectives. Fig. 4 presents normalized macroscopic density heatmaps of the ground truth and simulated results across three datasets. Representative frames at the early, middle, and late simulation stages are shown. SA-PINN consistently maintains density distributions closer to the ground truth throughout the simulation. Although initially minor, discrepancies from baseline simulations progressively increase, particularly in high-density (Cross 90) and spatially complex (GC) scenarios. In the Lane scenario with bidirectional pedestrian flows, SA-PINN effectively reproduces realistic stripe-like density patterns, capturing pedestrian lane formation phenomena. Fig. 3 shows the simulation snapshots of microscopic crowd behavior on the Lane dataset, in which each square shows pedestrian movements within localized regions at time steps ($t = 90, 100, 110$). Red and blue arrows represent pedestrian velocity directions, yellow circles highlight physically implausible behaviors (e.g., overlaps), and black ellipses mark the pedestrian following and formation-keeping behaviors observed. SA-PINN achieves simulation closely aligning with ground truth and exhibiting fewer physically implausible behaviors.

In addition, we compare the number of trainable parameters and training efficiency between SA-PINN and baseline methods. SA-PINN demonstrates clear advantages in both model compactness and computational efficiency, containing only 60K trainable parameters that are fewer than NSP (2.5M), PCS (0.6M), and SPDiff (0.2M). On the GC dataset, SA-PINN achieves a 25% reduction in training time compared to SPDiff, decreasing from 13.65 to 10.23 hours. Furthermore, training speed improvements of 23% and 11% are observed on the Cross 90 and Lane datasets, respectively.

4.3 Adaptivity Analysis

To demonstrate the adaptive capabilities of our method in crowd simulation, we visualize the attention weight distributions within the self-adaptive social perception module and track the evolving loss weights during multi-task training. These visualizations and corresponding analyses of intermediate representations and training dynamics underscore the interpretability of our model in modeling crowd perception and decision-making processes. Fig. 5 (a) and (b) show the Individual-attended influence maps under the scenarios of two typical behaviors (gap-seeking and following) in

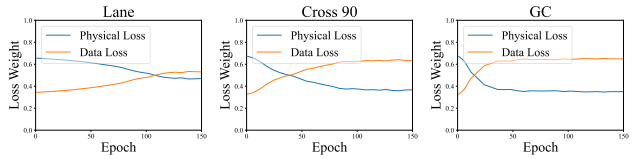


(a) Following Crowd

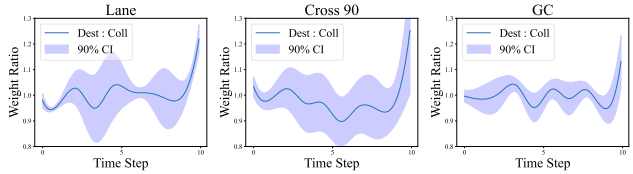
(b) Gap Seeking Crowd

Figure 5: Individual-attended influence maps: (a) the Following crowd attends to nearby agents in the same direction; (b) the Gap-Seeking crowd allocates attention toward forward gaps for navigation.

the Lane dataset. Blue dots indicate the focal pedestrian, red dots denote neighboring pedestrians, arrows represent velocities, and brighter areas show higher attention levels. Results demonstrate dynamic adaptation in pedestrian attention patterns: gap-seeking pedestrians primarily focus on open spaces ahead, while followers predominantly attend to the pedestrian directly in front.



(a) The physical and data loss weights evolve with the training epochs



(b) Destination to collision loss weight ratio evolves with time steps

Figure 6: Evolution of Multi-Task Loss Weights from TB-Net.

Fig. 6 (a) depicts adaptive changes in physical and data loss weights during training. Initially, physical losses dominate, guiding the model toward physically consistent behaviors. Subsequently, data losses gradually increase, enabling the learning of diverse individual behaviors. This balancing intersection occurs earlier for the larger GC dataset and later for the Lane dataset, reflecting the method's adaptive allocation of resources based on dataset characteristics. Fig. 6 (b) shows the relative weighting between goal loss and collision loss across time steps, highlighting a progressively greater emphasis on goal-reaching task as pedestrians approach their goals, aligning with realistic pedestrian behavior.

4.4 Ablation Study

We conduct ablation studies on key model components: the self-adaptive social perception model (SA-SP) and the task balancing network (TB-Net). Three variants are analyzed: (1) w/o SA-SP, where the SA-SP module is removed, leaving only individual motion features for prediction; (2) w/o TB-Net, where TB-Net is removed, relying solely on data loss; and (3) w/o both, removing both perception-level and task-level adaptive physics-informed components, resulting in a basic LSTM-based motion prediction network.

Table 2: Ablation study results for SA-PINN.

	Methods	Micro-metrics					Macro-metrics	
		MSE (↓)	MAE (↓)	OT (↓)	MMD (↓)	Col-(↓)	VD (↓)	DDS (↑)
Lane	w/o SA-SP	0.217	0.358	0.200	0.015	116	0.150	0.455
	w/o TB-Net	0.215	0.354	0.197	0.014	98	0.094	0.471
	w/o both	0.260	0.402	0.239	0.020	158	0.163	0.470
	SA-PINN	**0.165**	**0.291**	**0.152**	**0.011**	**62**	**0.089**	**0.474**
Cross 90	w/o SA-SP	0.563	0.505	0.293	0.030	204	0.125	0.327
	w/o TB-Net	0.562	0.505	0.291	0.028	192	0.114	0.346
	w/o both	0.773	0.668	0.593	0.090	190	0.100	0.276
	SA-PINN	**0.405**	**0.403**	**0.204**	**0.018**	**128**	**0.058**	**0.494**
GC	w/o SA-SP	2.290	1.088	2.178	0.030	398	0.268	0.478
	w/o TB-Net	2.145	1.106	1.947	0.041	162	0.330	0.440
	w/o both	6.228	1.875	3.948	0.068	708	0.245	0.346
	SA-PINN	**1.409**	**0.880**	**1.262**	**0.008**	**110**	**0.211**	**0.511**

Table 2 presents the experimental results for the aforementioned ablation study. While removing TB-Net leads to a certain degree of performance degradation, the model still outperforms all baselines on most datasets, except for a slight drop below NSP on the high-density Cross 90 dataset. This suggests that TB-Net plays an important role in handling high-density scenarios. Notably, the social perception attention network provides individuals with awareness and feedback regarding their surroundings. Removing SA-SP results in a more substantial performance decline, especially with a significant increase in collisions. This indicates that SA-SP contributes more critically to model performance, and the combination of TB-Net and SA-SP is particularly important for managing complex and high-density environments. Furthermore, after removing both TP-Net and SA-SP, the decline in microscopic metrics is more pronounced than in macroscopic metrics, emphasizing the importance of adaptive methods in enhancing the model's ability to learn diverse individual behavioral patterns from data.

5 Conclusion

In this study, we introduced a Self-Adaptive Physics-Informed Neural Network (SA-PINN) architecture for crowd simulation, aimed at modeling the dynamic nature of pedestrian perception and motion planning. We developed a self-adaptive social perception model to capture evolving social attention patterns among pedestrians. Furthermore, we proposed a self-adaptive multi-task PINN training framework that improves both the efficiency and effectiveness of motion prediction by incorporating a physics-informed multi-task loss with dynamically adjusted weights. Extensive experiments across diverse scenarios show that our model consistently outperforms recent state-of-the-art PINN-based approaches and generates more reliable and physically plausible crowd behaviors. In future work, we will further explore the potential of multimodal data to enhance crowd behavior understanding. For example, we plan to incorporate semantic maps or leverage vision-language models (VLMs) to parse synthetic simulation videos and extract textual representations of crowd rules, thereby enriching the input data modalities. This line of research will facilitate the modeling of complex crowd dynamics in both normal and emergency situations.

Acknowledgments

This work is supported in part by the Key Research and Development Program of Shaanxi under Program 2025GH-YBXM-020, in part by the 111 Center under Grant B16037, in part by the National Research Foundation (NRF), Singapore, through the AI Singapore Programme under the project titled "AI-based Urban Cooling Technology Development" (Award No. AISG3-TC-2024-014-SGKR), and in part by the Fundamental Research Funds for the Central Universities (No. YJSJ25011).

References

[1] Karol A Bacik, Bogdan S Bacik, and Tim Rogers. 2023. Lane nucleation in complex active flows. *Science* 379, 6635 (2023), 923–928.

[2] Maik Boltes, Jun Zhang, and Armin Seyfried. 2013. *Analysis of Crowd Dynamics with Laboratory Experiments.* Springer New York.

[3] Edwin V Bonilla, Kian Chai, and Christopher Williams. 2007. Multi-task Gaussian process prediction. *Advances in Neural Information Processing Systems* 20 (2007).

[4] Panayiotis Charalambous, Julien Pettre, Vassilis Vassiliades, Yiorgos Chrysanthou, and Nuria Pelechano. 2023. Greil-crowds: Crowd simulation with deep reinforcement learning and examples. *ACM Transactions on Graphics* 42, 4 (2023), 1–15.

[5] Hongyi Chen, Jingtao Ding, Yong Li, Yue Wang, and Xiao-Ping Zhang. 2024. Social physics informed diffusion model for crowd simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 474–482.

[6] Mengyu Chu, Lingjie Liu, Quan Zheng, Erik Franz, Hans-Peter Seidel, Christian Theobalt, and Rhaleb Zayer. 2022. Physics informed neural fields for smoke reconstruction with sparse data. *ACM Transactions on Graphics* 41, 4 (2022), 1–14.

[7] Emilio Jose Rocha Coutinho, Marcelo Dall'Aqua, Levi McClenny, Ming Zhong, Ulisses Braga-Neto, and Eduardo Gildin. 2023. Physics-informed neural networks with adaptive localized artificial viscosity. *J. Comput. Phys.* 489 (2023), 112265.

[8] Dirk Helbing and Peter Molnar. 1995. Social force model for pedestrian dynamics. *Physical review E* 51, 5 (1995), 4282.

[9] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. 2022. Neural mocon: Neural motion control for physically plausible human motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6417–6426.

[10] Reiya Itatani and Nuria Pelechano. 2024. Social Crowd Simulation: Improving Realism with Social Rules and Gaze Behavior. In *Proceedings of the 17th ACM SIGGRAPH Conference on Motion, Interaction, and Games*. 1–11.

[11] Yuling Jiao, Di Li, Xiliang Lu, Jerry Zhijian Yang, and Cheng Yuan. 2024. A Gaussian mixture distribution-based adaptive sampling method for physics-informed neural networks. *Engineering Applications of Artificial Intelligence* 135 (2024), 108770.

[12] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. Physics-informed machine learning. *Nature Reviews Physics* 3, 6 (2021), 422–440.

[13] Hoshang Kolivand, Mohd Shafry Rahim, Mohd Shahrizal Sunar, Ahmad Zakwan Azizul Fata, and Chris Wren. 2021. An integration of enhanced social force and crowd control models for high-density crowd simulation. *Neural Computing and Applications* 33 (2021), 6095–6117.

[14] Xiao-Cheng Liao, Wei-Neng Chen, Xiao-Qi Guo, Jinghui Zhong, and Xiao-Min Hu. 2023. Crowd management through optimal layout of fences: An ant colony approach based on crowd simulation. *IEEE Transactions on Intelligent Transportation Systems* 24, 9 (2023), 9137–9149.

[15] Linbo Luo, Cheng Chai, Jianfeng Ma, Suiping Zhou, and Wentong Cai. 2018. ProactiveCrowd: Modelling Proactive Steering Behaviours for Agent-Based Crowd Simulation. *Computer Graphics Forum* 37, 1 (2018), 375–388.

[16] Linbo Luo, Baodan Zhang, Bin Guo, Jinghui Zhong, and Wentong Cai. 2022. Why they escape: Mining prioritized fuzzy decision rule in crowd evacuation. *IEEE Transactions on Intelligent Transportation Systems* 23, 10 (2022), 19456–19470.

[17] Zhaobin Mo, Yongjie Fu, and Xuan Di. 2024. PI-NeuGODE: Physics-Informed Graph Neural Ordinary Differential Equations for Spatiotemporal Trajectory Prediction. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 1418–1426.

[18] Mehdi Moussaïd, Dirk Helbing, and Guy Theraulaz. 2011. How simple rules determine pedestrian behavior and crowd disasters. *Proceedings of the National Academy of Sciences* 108, 17 (2011), 6884–6888.

[19] Renbin Pan, Feng Xiao, and Minyu Shen. 2024. ro-PINN: A reduced order physics-informed neural network for solving the macroscopic model of pedestrian flows. *Transportation Research Part C: Emerging Technologies* 163 (2024), 104658.

[20] Andreas Panayiotou, Theodoros Kyriakou, Marilena Lemonari, Yiorgos Chrysanthou, and Panayiotis Charalambous. 2022. CCP: Configurable Crowd Profiles. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference*. 53:1–53:10.

[21] Jiaping Ren, Wei Xiang, Yangxi Xiao, Ruigang Yang, Dinesh Manocha, and Xiaogang Jin. 2019. Heter-sim: Heterogeneous multi-agent systems simulation by interactive data-driven optimization. *IEEE Transactions on Visualization and Computer Graphics* 27, 3 (2019), 1953–1966.

[22] Liangchen Song, Sheng Liu, Celong Liu, Zhong Li, Yuqi Ding, Yi Xu, and Junsong Yuan. 2021. Learning kinematic formulas from multiple view videos. In *Proceedings of the 29th ACM International Conference on Multimedia*. 126–134.

[23] Jur Van Den Berg, Stephen J Guy, Ming Lin, and Dinesh Manocha. 2011. Reciprocal n-body collision avoidance. In *Robotics Research - The 14th International Symposium ISRR*. 3–19.

[24] Wouter van Toll, Thomas Chatagnon, Cédric Braga, Barbara Solenthaler, and Julien Pettré. 2021. SPH crowds: Agent-based crowd simulation up to extreme densities using fluid dynamics. *Computers & Graphics* 98 (2021), 306–321.

[25] Jian Cheng Wong, Chin Chun Ooi, Abhishek Gupta, and Yew-Soon Ong. 2022. Learning in sinusoidal spaces with physics-informed neural networks. *IEEE Transactions on Artificial Intelligence* 5, 3 (2022), 985–1000.

[26] Hao Wu, Fan Xu, Chong Chen, Xian-Sheng Hua, Xiao Luo, and Haixin Wang. 2024. Pastnet: Introducing physical inductive biases for spatio-temporal video prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2917–2926.

[27] Wenhan Wu, Maoyin Chen, Jinghai Li, Binglu Liu, Xiaolu Wang, and Xiaoping Zheng. 2022. Visual information based social force model for crowd evacuation. *Tsinghua Science and Technology* 27, 3 (2022), 619–629.

[28] Wenfeng Yi, Wenhan Wu, Xiaolu Wang, Erhui Wang, and Xiaoping Zheng. 2024. Order–disorder phase transitions in front of the exit during human crowd evacuations. *Transportation Research Part C: Emerging Technologies* 163 (2024), 104649.

[29] Jiangbei Yue, Dinesh Manocha, and He Wang. 2022. Human trajectory prediction via neural social physics. In *Proceedings of the 17th European Conference on Computer Vision*. 376–394.

[30] Bosi Zhang, Youmei Gao, Yong Han, Siyi Liang, Qiaolin Chen, and Zhihong Yu. 2022. Walking characteristics and collision avoidance strategy in bidirectional pedestrian flow: A study focused on the influence of social groups. *Journal of Statistical Mechanics: Theory and Experiment* 2022, 7 (2022), 073405.

[31] Guozhen Zhang, Zihan Yu, Depeng Jin, and Yong Li. 2022. Physics-infused machine learning for crowd simulation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2439–2449.

[32] Sainan Zhang, Jun Zhang, Mohcine Chraibi, and Weiguo Song. 2021. A speed-based model for crowd simulation considering walking preferences. *Communications in Nonlinear Science and Numerical Simulation* 95 (2021), 105624.

[33] Zhibo Zhang, Yanjun Zhu, Rahul Rai, and David Doermann. 2022. Pimnet: Physics-infused neural network for human motion prediction. *IEEE Robotics and Automation Letters* 7, 4 (2022), 8949–8955.

[34] Jinghui Zhong, Dongrui Li, Zhixing Huang, Chengyu Lu, and Wentong Cai. 2022. Data-driven crowd modeling techniques: A survey. *ACM Transactions on Modeling and Computer Simulation* 32, 1 (2022), 1–33.

[35] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. 2012. 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*. 2871–2878.

[36] Yanshan Zhou, Pingrui Lai, Jiaqi Yu, Yingjie Xiong, and Hua Yang. 2024. Hydrodynamics-Informed Neural Network for Simulating Dense Crowd Motion Patterns. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 4553–4561.