

Continuous Locomotive Crowd Behavior Generation

Inhwan Bae, Junoh Lee and Hae-Gon Jeon*

Gwangju Institute of Science and Technology, South Korea

{inhwanbae, juno}@gm.gist.ac.kr, haegonj@gist.ac.kr

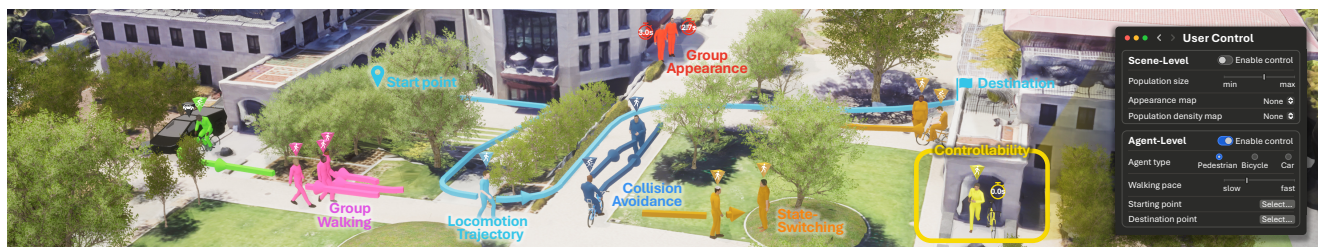


Figure 1. Generating realistic, continuous crowd behaviors with learned dynamics. Given a scene image, CrowdES iteratively populates the environment and synthesizes diverse locomotion patterns to create lifelike crowd scenarios. CrowdES also allows users to control parameters to achieve tailored and flexible outcomes.

Abstract

Modeling and reproducing crowd behaviors are important in various domains including psychology, robotics, transport engineering and virtual environments. Conventional methods have focused on synthesizing momentary scenes, which have difficulty in replicating the continuous nature of real-world crowds. In this paper, we introduce a novel method for automatically generating continuous, realistic crowd trajectories with heterogeneous behaviors and interactions among individuals. We first design a crowd emitter model. To do this, we obtain spatial layouts from single input images, including a segmentation map, appearance map, population density map and population probability, prior to crowd generation. The emitter then continually places individuals on the timeline by assigning independent behavior characteristics such as agents' type, pace, and start/end positions using diffusion models. Next, our crowd simulator produces their long-term locomotions. To simulate diverse actions, it can augment their behaviors based on a Markov chain. As a result, our overall framework populates the scenes with heterogeneous crowd behaviors by alternating between the proposed emitter and simulator. Note that all the components in the proposed framework are user-controllable. Lastly, we propose a benchmark protocol to evaluate the realism and quality of the generated crowds in terms of the scene-level population dynamics and the individual-level trajectory accuracy. We demonstrate that our approach effectively models diverse crowd behavior patterns and generalizes well across different geographical environments. Code is publicly available at <https://github.com/InhwanBae/CrowdES>.

*Corresponding author

1. Introduction

Crowds exhibit highly complex behaviors, because people are driven by individual goals, and are influenced by other crowd members and environmental factors [196]. Synthesizing these behaviors is crucial for various applications such as autonomous driving [60], computer games [36], virtual cinematography [135] and urban planning [168]. Unfortunately, creating vivid, lifelike crowd behaviors requires labor-intensive human annotation. Artists or graphic designers must manually position agents in virtual environments, and employ handcrafted intelligence to replicate natural behavioral motions.

To accelerate these tasks, many efforts have been made to automatically populate the environments using interactive authoring [169] and procedural modeling [72, 199], and motion synthesis using rule-[135]/force-[50]/velocity-based approaches [170]. Learning-based methods have recently been used to facilitate realistic behavior planning by integrating visual information and contextual cues [19, 71, 114, 132]. However, because a wide variety of crowd collective patterns emerge from self-organization, as yet a holistic framework for learning and generating realistic human dynamics is, to the best of our knowledge, still unavailable.

Meanwhile, research into automatic actor placement has begun for traffic scene generation [31, 38, 161]. In this task, generative models augment plausible agent layouts on high-definition maps of the surrounding area [93, 128, 160]. However, by generating momentary 2D distributions of vehicles, such methods are hardly able to handle the continuous, evolving nature of real-world environments.

In this paper, our goal is to automatically produce con-

tinuous and realistic crowd trajectories with heterogeneous behaviors and interactions in space from single input images. To achieve this, we introduce CrowdES, a framework consisting of two models: the crowd emitter and simulator.

The crowd emitter, inspired by particle systems [16], generates individuals who are characterized by various attributes, and continuously populates the scene over time. The emitter firstly analyzes the input image and then predicts spatial layouts, including a semantic segmentation map, appearance map, population density map and population probability. Here, the appearance map indicates potential locations where people might appear, (*e.g.*, building entrances, scene boundaries) and the population map assesses the crowded areas where people are likely to gather. The population probability computes categorical distribution based on the number of agents in the scene. The number of agents to appear within a certain time-duration are determined based on the distribution. Next, the emitter generates new agents conditioned on the current crowd distribution and the predicted spatial layout. For this, a diffusion model is adopted to produce parameters for each agent, including the agents' attributes, starting and destination coordinates, paces, and appearance times. This allows the timeline of each agent to be planned, where individuals appear at their starting points within a specific time window.

The crowd simulator, inspired by locomotion simulation, produces the trajectory from each agent's starting to end-points. Unlike conventional methods, our simulator enables intermediate behavior augmentations, such as stopping to chat with other agents and avoiding obstacles by moving left or right. Motivated by the Switching Dynamical System (SDS), we define various behavior patterns of the crowd as behavioral modes. Using encoded features for the agent attributes, past behaviors and interactions with neighboring individuals, our simulator computes the transition probabilities between the behavioral modes. The mode is randomly sampled under a Markov chain and fed into the decoder to make decisions about the agent's next steps. Our crowd simulator generates the full route toward the destination through the recurrent prediction process. Once an agent reaches its destination, it disappears from the scene.

Our CrowdES can produce lifelong crowd animations by alternating between the crowd emitter and simulator. Additionally, it allows users to control intermediate outputs with flexibility to customize the population density, starting and destination positions, actor types and walking speed.

Lastly, we build a **benchmark protocol** to evaluate the performance of our CrowdES and relevant works. Inspired by trajectory prediction tasks, we compare the generated result to ground-truth crowd tracks considering two aspects: (1) assessing realism at the scene level based on crowd distribution and (2) evaluating the accuracy of individual agent trajectories. This benchmark is performed with up to 10

hours of video. Our framework, which benefits from the synergy between the emitter and simulator, can generate realistic crowd animations, even in previously unseen environments. Furthermore, the flexibility of our framework allows us to achieve controllable crowd scenarios.

2. Related Works

2.1. Multi-Agent Trajectory Prediction

Starting with physical formulation-based methods [50, 104, 120, 194], trajectory prediction models have significantly improved by incorporating neural networks and learning techniques. To infer socially-acceptable paths, many considerations have been implemented for both the agents' interaction and dynamics modeling. Interactions with neighbors are crucial, especially to adhere to social norms such as collision avoidance and group movements. One pioneering work, Social-LSTM [2], implicitly models the social relations of agents by integrating hidden states for their neighbors through a social pooling module. Subsequent research weighs the mutual influences among agents using attention mechanisms [40, 55, 142, 175], graph convolutional networks [5, 20, 63, 64, 90, 105, 106, 156], graph attention networks [8, 51, 56, 84, 85, 147, 163, 174, 180, 187], and transformers [7, 42, 108, 126, 148, 167, 181–183, 189, 201, 202]. Incorporating additional visual data allows the trajectory prediction models to leverage semantic information about the walkable terrain [4, 17, 23, 25, 29, 37, 39, 58, 65, 83, 98–102, 119, 125, 127, 140, 144, 158, 164, 165, 173, 177, 178, 193, 195, 203, 204, 209, 211].

The captured interaction features are then used for dynamics models to predict feasible future trajectories. Predictors adopt either recurrent methods [2, 15, 22, 26, 27, 32, 94, 95, 103, 111, 117, 122, 130, 131, 190, 192, 205–208], which account for the temporal characteristics of trajectory coordinates, or simultaneous methods [5, 6, 43, 52, 76, 81, 88, 105, 115, 145, 186], which regress all coordinates at once. Recent works have introduced probabilistic inferences to explain the indeterminacy inherent in crowd behaviors [47]. Techniques such as bivariate Gaussian distributions [5, 80, 105, 109, 145, 149, 150, 191, 197], generative adversarial networks [28, 47, 51, 82, 86, 155, 159], conditional variational autoencoders [14, 21, 74, 75, 79, 97, 116, 118, 157, 176, 185, 188], diffusion models [46, 57, 100, 132, 179], language models [10, 123, 143], and explicit modeling [9, 11, 67, 146] have been used for stochastic trajectory prediction.

Although these trajectory prediction models can generate realistic and diverse behaviors, they face several challenges when attempting unconditional generation, because of their dependency on past trajectories. Additionally, because they use fixed time windows for both the observation and inference steps, it is difficult to plan long-term paths.

2.2. Crowd Locomotion Simulation

Boids algorithms [135, 136], one of the earliest crowd simulation systems, introduce simple rules for alignment, cohesion, and separation to model group dynamics. Since then, crowd simulations have been studied to leverage various group behaviors [134], for collision avoidance [170, 171], and user-defined constraints and goals [35, 48, 62, 121]. Data-driven methods, which first create databases of example behaviors and then match [73, 77] or blend [59, 78] agents' actions to them during simulations, have further enhanced the diversity and realism of output actions [69, 70, 210]. These approaches have evolved to the development of learnable approaches [24, 71, 114, 132, 198]. For instance, GREIL-Crowds [19] trains a goal-seeking behavior within a group using guided reinforcement learning. Additionally, it shows continuous crowd simulations by leveraging real-world data, including individual entrance times, origins and goals. However, research on methods to continuously populate scenes with crowds beyond just locomotion, remains limited.

Populating spaces with crowds that exhibit natural behaviors is also critical for broader applications. Commercial software for 3D animation [151], visual effects [45], urban planning [168] and games [36] employ functions which treat virtual crowds as particles, in which particle emitter systems randomly generate actors at predefined regions. In this paper, we shift this paradigm into a learnable method to plan how densely to populate scenes with crowds and perform crowd emissions using diffusion models.

2.3. Traffic Scene Generation

Recent advances in generative models have enabled the synthesis of realistic traffic scenes [30, 60, 93, 161, 162]. This task generates the initial position, direction and size of vehicles in a scene, given a map image. Specifically, SceneGen [161] employs an LSTM module to autoregressively generate traffic scenes by sequentially inserting actors one at a time. TrafficGen [38] uses an encoder-decoder architecture to sample vehicles' initial states in probability distributions. RealGen [31] synthesizes traffic by fusing retrieved examples from external data. More recently, several methods have been proposed to leverage the powerful generative abilities of diffusion models for vehicle placement [93, 128, 129, 160]. Although these works have demonstrated the potential ability to learn to populate environments, they still only focus on agent placement during the initial scene setup. As a result, it is difficult to account for vehicles that enter later. So, once all of the vehicles have left the area of interest, the scenes become empty.

In this paper, we present a diffusion model-based crowd emitter which continually populates scenes with dynamic actors, ensuring lifelong crowd animations.

2.4. Switching Dynamical Systems

A dynamical system is a framework based on a set of rules or equations [13, 138]. For complex dynamical systems, dividing their behaviors into distinct modes, each with simpler dynamics, is often effective [87]. Switching Dynamical Systems (SDS) facilitates the identification of these modes and the transitions between them on time series data [1, 12, 41, 44, 68, 89, 112]. In particular, SNLDS [33] learns to switch between the discrete states of nonlinear dynamical models. REDSDS [3] introduces a recurrent state-to-switch connection, along with explicit state duration models, to efficiently capture the duration of varying states. GRASS [91] further advances this approach by employing a dynamic graph-based aggregation to model interaction-aware mode switching. We incorporate SDS into crowd dynamics modeling to better generate agents' dynamic, long-term movement and the behaviors of crowds.

3. Methodology

We describe a method for modeling the continuous dynamics of crowd behaviors. First, we define the crowd behavior generation problem in Sec. 3.1. Next, in Sec. 3.2, we introduce our crowd emitter model to populate environments made from single input images. We then present our crowd simulator model, which produces trajectories from the starting points to the destinations of each agent with intermediate behavior augmentations, in Sec. 3.3. Finally, in Sec. 3.4, we integrate both the crowd emitter and crowd simulator within the proposed CrowdES framework. An overview of our CrowdES framework is illustrated in Fig. 2.

3.1. Problem Definition

Our goal is to predict realistic trajectories of crowds over time based on single input images. Specifically, given a scene image $\mathcal{I}$, we aim to generate a crowd behavior scenario $\mathcal{V}$ of length $T_{\mathcal{V}}$ with N agents in total. Each agent $A = \{\kappa, \mathcal{T}\}$ is characterized by an agent type κ and a trajectory $\mathcal{T} = [c_{T_s}, \dots, c_{T_d}]$, where c_t represents the 2D coordinate (x, y) at time t . Here, T_s and T_d denote the start and destination times for each agent, respectively.

Modeling long and multiple trajectories is challenging because future crowd behaviors are highly correlated with past scene states and interactions with each other. To alleviate this complexity, we adopt an approach that generates crowd behaviors incrementally within a smaller time window of length T_w . Within each time window, we capture the emerging and locomotion characteristics of individual agents using a two-stage modeling approach: the crowd emitter and the crowd simulator.

3.2. Crowd Emitter Model

To understand terrain geometry and to ensure the controllability of the crowd emitter during inference, we design a pre-

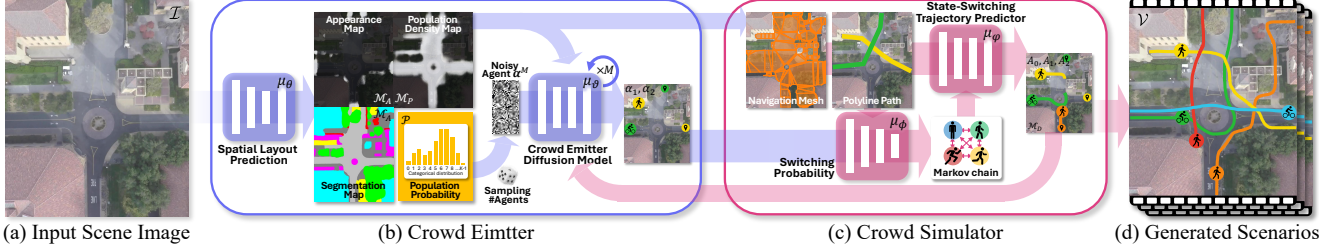


Figure 2. An overview of our CrowdES framework. Starting with the input scene image $\mathcal{I}$, CrowdES continuously generates realistic crowd behaviors $\mathcal{V}$ by alternating between the crowd emitter and crowd simulator processes.

processing step that assesses spatial layouts from input scene images. The diffusion model for generating populations then uses the spatial layouts as conditions in a time-sequential manner.

Spatial layout prediction. We extract a spatial layout from input scene images $\mathcal{I}$, consisting of a semantic segmentation map $\mathcal{M}_S$, an appearance map $\mathcal{M}_A$, a population density map $\mathcal{M}_P$, and population probability $\mathcal{P}$.

The semantic segmentation maps $\mathcal{M}_S$ categorize each pixel of the scenes into seven classes: buildings, structures, bushes, grasses, trees, sidewalks and roads. The appearance maps $\mathcal{M}_A$ highlight possible locations for people to emerge. To establish the ground-truth of $\mathcal{M}_A$ in a binary image, we gather coordinates c_{T_s} and c_{T_d} for all N agents, then mark the pixels as 1. The population density maps identify regions where people are likely to come together. We create the $\mathcal{M}_P$ label by counting the number of pixels on a trajectory coordinate τ for all N agents across the image pixels, by applying a logarithmic transformation, and by normalizing the counting values between 0 and 1. The population probabilities $\mathcal{P}$ is a probability made up of K numbers for 0 to $K - 1$ population, and represent a categorical distribution of how many agents will exist in the current scene. The ground-truth labels for $\mathcal{P}$ are computed by counting the number of agents present across all T_V frames.

We compute $\mathcal{M}_S$ using a pretrained Grounded-SAM [133] model. We also leverage a pretrained SegFormer [184] backbone μ_θ to obtain the others, a segmentation head to predict $\mathcal{M}_A$ & $\mathcal{M}_P$, and a classification head to estimate $\mathcal{P}$ as:

$$\mathcal{M}_A, \mathcal{M}_P, \mathcal{P} = \mu_\theta(\mathcal{I}, \mathcal{M}_S). \quad (1)$$

Crowd emitter diffusion model. To populate the scenes, we need to determine the number of agents $N_{t:t+T_w}$ to be assigned over the interval t to $t + T_w$. This value $N_{t:t+T_w}$ is sampled based on the number of agents in the previous frame N_{t-1} , and the population probability distribution $\mathcal{P}$, defined by the following equation:

$$N_{t:t+T_w} = \min(N_{\mathcal{P}} - N_{t-1}, 0) \quad (2)$$

where $N_{\mathcal{P}} \sim \text{Categorical}(\mathcal{P})$.

Next, we implement a diffusion model to iteratively generate agents through a denoising process. Each agent is

parameterized by $\alpha = \{\kappa, \nu, T_s, c_{T_s}, c_{T_d}\}$, where ν denotes the agents' walking pace. The final agent parameters α^0 are progressively denoised from an initial Gaussian noise vector α^M over a sequence of M diffusion steps. In the forward diffusion process, noise is gradually added to the data as:

$$q(\alpha^{1:M} | \alpha^0) := \prod_{m=1}^M q(\alpha^m | \alpha^{m-1}) \quad (3)$$

$$q(\alpha^m | \alpha^{m-1}) := \mathcal{N}(\alpha^m; \sqrt{1 - \beta^m} \alpha^{m-1}, \beta^m \mathbf{I}),$$

where β_m is a constant that defines the noise schedule at each step m . The reverse denoising process restores α^M to α^0 using a learnable network μ as follows:

$$p_\theta(\alpha^{0:M}) := p(\alpha^M) \prod_{m=1}^M p_\theta(\alpha^{m-1} | \alpha^m), \quad (4)$$

$$p_\theta(\alpha^{m-1} | \alpha^m) := \mathcal{N}(\alpha^{m-1}; \mu_\theta(\alpha^m, m, \mathbf{C}_e), \beta^m \mathbf{I}).$$

Here, $\mathbf{C}_e = \{\mathcal{M}_S, \mathcal{M}_A, \mathcal{M}_P, \mathcal{M}_D\}$ represents the environmental conditions guiding the agent parameters to align with terrain constraints. Specifically, $\mathcal{M}_D$ is a binary map indicating individual positions in the previous frame. The diffusion emitter generates clean agent parameters from Gaussian noise by using a reverse denoising process.

In addition, individuals in crowds often form groups, move collectively, and share their destinations [110]. To model this collectivity, our model denoises the parameters of all $N_{t:t+T_w}$ agents, $[\alpha_1, \dots, \alpha_{N_{t:t+T_w}}]$, simultaneously. To be specific, the denoising network μ_θ leverages transformer architectures to propagate features across agents, and then uses cross-attention with encoded environmental conditions to improve the spatial awareness. In the end, the generated agents are positioned along a timeline, and are placed in the scene by the crowd simulator.

3.3. Crowd Simulator Model

To generate environment-aware locomotion that allows agents to navigate from their starting points to destinations, even across complex terrains, we employ a navigation mesh and its benefits. Additionally, we design a network to predict agent behavior dynamics, enabling stochastic switching to achieve continuous and diverse crowd behaviors.

Navigation mesh. In simple and open spaces, paths to

destinations can often be planned directly with minimal complexity. However, in more complex environments, such as a university campus, routes may need to navigate around various obstacles, sometimes requiring U-shaped or maze-like paths. To simulate both cases, we use a navigation mesh, subdividing the walkable spaces into polygonal regions for path planning [152, 166, 172], to define an initial path to the destination.

First, we create a binary traversable map $\mathcal{M}_W$ by using the segmentation map $\mathcal{M}_S$ to exclude non-navigable areas such as buildings, structures, and bushes. We then exploit the Recast method [54, 107] to generate a traversable mesh. Using the connectivity of polygons within a mesh, we are able to search for a polyline path from the agent A 's current position c_t to the destination c_{T_d} . Along this polyline, we designate $c_{t,nav}$, a control point located at a distance equal to the agent's walking pace ν_n from c_t , for navigation.

Switching crowd dynamical systems. For state-switching behavior, we predict the transition probabilities of a Markov chain based on historical data. Inspired by anchor-based trajectory prediction methods [9, 18, 67], we define the behavior states in a data-driven manner. Each agent's trajectory $\mathcal{T}$ is segmented into sequences of length T_f , yielding $\mathcal{T}_{t:t+T_f}$. The parameter T_f determines how frequently agents can take different behaviors. Using the origin coordinate c_t of each segmented trajectory, we calculate $c_{t,nav}$. We then normalize the segmented trajectory in terms of translation, rotation, and scale using $c_{t,nav}$. Here, we apply K-means clustering to obtain B cluster centers, each representing one of the B behavioral states. The index of the cluster center closest to $\mathcal{T}_{t:t+T_f}$ is assigned as the ground-truth behavior state $\bar{b}_f$.

Next, we predict an agent's future state b_f based on its previous state b_h . In particular, the learnable network μ_ϕ takes agent parameters, historical trajectory, neighborhood trajectories and environmental information as conditions $\mathcal{C}_s$ to predict b_f as follows:

$$p_\phi(b_f|b_h) := \text{Categorical}(b_f; \mu_\phi(b_h, \mathcal{C}_s), b_h) \quad (5)$$

$$\text{s.t. } \mathcal{C}_s = \{\alpha, c_{t,nav}, \mathcal{T}_{t-T_h:t}, \mathcal{H}_{t-T_h:t}, \mathcal{M}_S, \mathcal{M}_P\},$$

where $\mathcal{T}_{t-T_h:t}$ is the historical trajectory of length T_h , and $\mathcal{H}_{t-T_h:t} = \{\mathcal{T}_{t-T_h:t}\}_{n=1}^{N_{t-1}}$ is the neighborhood trajectories used to capture social interactions. The sampled behavior state is then fed into the trajectory prediction network.

State-switching crowd simulator. Lastly, we generate agent locomotion in a recurrent manner. Using the agent-environment feature set $\mathcal{C}_s$ and a sampled behavior state b_f , our trajectory predictor μ_ϕ estimates a realistic future trajectory as follows:

$$\mathcal{T}_{t:t+T_f} = \mu_\phi(b_f, \mathcal{C}_s). \quad (6)$$

Through an iterative inference with μ_ϕ , each agent can have a complete sequence of footsteps from the starting point to the destination.

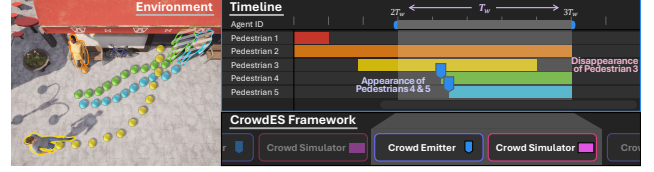


Figure 3. An overview of our recurrent crowd behavior generation approach using the crowd emitter and simulator. (Marker: Emerging crowds at specific times, Bar: Extensible crowd trajectories.)

During network training, we use a ground-truth behavior $\bar{b}_f$, instead of the sampled b_f . This enables our model to implement a dynamical system where crowd dynamics can be changed at intervals of T_f .

3.4. Continuous Crowd Behavior Generation

CrowdES framework. Finally, we incorporate the proposed emitter and simulator into our CrowdES framework to generate infinitely long crowd animations over time. For this, we iteratively generate crowd behaviors with a length T_w by alternating between the crowd emitter and simulator, as illustrated in Fig. 3. The crowd emitter places the agent markers on the timeline to indicate when agents will emerge over the next T_w frames. The crowd simulator then generates trajectories within this window for all agents in the scene. By repeating this emission-simulation process, our CrowdES produces continuous crowd behaviors, given the environment.

In real-world video footage, people obviously exist throughout the entire space. However, agents in our framework appear one-by-one at the first iteration. To address this, we employ a simple trick: we only populate a half of the total agents $N_{0:T_w}$ during the iteration from 0 to $T_w - 1$, and use a frame at the iteration T_w as an initial frame.

Implementation details. We empirically set T_w to 50 frames with a 5 fps system. For the crowd emitter model, we use $M = 50$ diffusion steps and a DDIM scheduler [154] for training. To train μ_θ , we use a binary cross entropy (BCE) as a loss function between the predicted logits with ground truth maps and probabilities. For μ_ψ , we employ a mean square error (MSE) between the output and Gaussian noise.

In the crowd simulation, we set $B = 8$ behavior states, $T_h = 10$, and $T_f = 20$ in order to prevent frequent behavior switching and ensure smooth motion transitions during simulation. For agent-centric predictions, we crop a map with 16×16 meters centered on each agent's coordinate. To train μ_ϕ , we exploit a cross entropy (CE) loss between the estimated and its ground truth behavior states. For μ_ψ , we use a mean absolute error (MAE) as a loss function between the predicted trajectory and the actual data.

The training is performed with the AdamW optimizer [92] at a learning rate of 0.0001. The batch size/training epochs are set to 512/256 and 2048/64 for the crowd emitter and the crowd simulator, respectively. All experiments are con-

Dataset	Model	Scene-Level Realism				Agent-Level Accuracy			
		Dens.	Freq.	Cov.	Pop.	Kinem.	DTW	Div.	Col.
ETH	SE-ORCA	0.054	0.034	0.034	0.640	<u>0.502</u>	<u>2.122</u>	0.188	0.054
	VAE	<u>0.032</u>	<u>0.025</u>	<u>0.025</u>	<u>0.351</u>	1.621	3.619	0.285	4.959
	CrowdES	0.020	0.020	0.020	0.208	0.377	1.649	<u>0.203</u>	<u>0.697</u>
HOTEL	SE-ORCA	<u>0.083</u>	<u>0.065</u>	<u>0.065</u>	0.746	0.458	<u>0.648</u>	0.248	0.045
	VAE	0.151	0.110	0.110	1.353	0.315	0.780	0.217	1.807
	CrowdES	0.013	0.009	0.009	0.117	<u>0.336</u>	0.643	<u>0.242</u>	<u>1.197</u>
UNIV	SE-ORCA	0.425	0.273	0.273	0.901	0.529	<u>1.279</u>	0.334	0.011
	VAE	<u>0.416</u>	<u>0.263</u>	<u>0.263</u>	<u>0.879</u>	<u>0.465</u>	1.332	0.402	0.460
	CrowdES	0.347	0.204	0.204	0.734	0.420	1.121	<u>0.340</u>	0.645
ZARA1	SE-ORCA	0.025	0.018	0.018	0.389	<u>0.473</u>	<u>1.256</u>	<u>0.227</u>	0.014
	VAE	0.014	0.008	0.008	0.231	0.576	1.342	0.221	<u>0.827</u>
	CrowdES	<u>0.018</u>	<u>0.017</u>	<u>0.017</u>	<u>0.254</u>	0.327	0.675	0.304	1.018
ZARA2	SE-ORCA	0.063	0.039	0.039	0.674	0.515	<u>1.244</u>	0.218	0.000
	VAE	<u>0.049</u>	<u>0.026</u>	<u>0.026</u>	<u>0.534</u>	<u>0.473</u>	1.309	<u>0.244</u>	<u>0.630</u>
	CrowdES	0.009	0.013	0.013	0.100	0.227	0.579	0.355	1.021
SDD	SE-ORCA	0.058	<u>0.047</u>	<u>0.044</u>	<u>0.540</u>	<u>0.893</u>	<u>6.645</u>	<u>0.377</u>	0.048
	VAE	<u>0.052</u>	0.052	0.047	0.678	1.078	7.881	0.387	2.470
	CrowdES	0.038	0.033	0.030	0.463	0.650	6.352	0.354	<u>0.411</u>
GCS	SE-ORCA	1.441	0.066	0.066	0.617	<u>3.419</u>	<u>3.289</u>	0.197	0.316
	VAE	<u>0.826</u>	0.085	0.085	<u>0.433</u>	5.529	5.351	<u>0.227</u>	9.025
	CrowdES	0.584	<u>0.066</u>	<u>0.066</u>	0.329	1.341	3.864	0.279	<u>8.519</u>
EDIN	SE-ORCA	<u>0.016</u>	<u>0.015</u>	<u>0.015</u>	<u>0.522</u>	<u>2.142</u>	<u>1.819</u>	0.221	0.000
	VAE	0.031	0.031	0.031	1.375	4.704	2.292	<u>0.375</u>	0.001
	CrowdES	0.002	0.002	0.002	0.313	0.471	1.361	0.386	0.000

Table 1. Comparison of the CrowdES framework with algorithmic and learnable methods. For Div., higher values indicate better performance; for all other metrics, lower values are better. **Best**, Underline: Second best.

ducted on an NVIDIA RTX 4090 GPU, typically taking about one day for each model.

4. Experiments

In this section, we conduct comprehensive experiments to verify the effectiveness of our CrowdES framework. We first describe our benchmark setup in Sec. 4.1. Next, we present comparison results with various baseline models on real-world scenes in Sec. 4.2. We then assess the flexibility and controllability of our framework in both challenging real-world and synthetic environments in Sec. 4.3. Finally, we perform extensive ablation studies to demonstrate the effects of each component of our framework in Sec. 4.4.

4.1. Benchmark Method

Datasets. To evaluate the realism of the generated crowd scenarios, we used five datasets, including ETH [120], UCY [77], Stanford Drone Dataset (SDD) [137], Grand Central Station dataset (GCS) [200, 212], and Edinburgh dataset (EDIN) [96].

The ETH-UCY datasets consist of five subsets, containing a total of 2,329 pedestrians recorded for over one hour of surveillance video. We re-label the dataset to address issues with fragmented and missing trajectory segments. Following [47], we adopt a leave-one-out strategy for both training and evaluation. The SDD dataset includes 10,065 agents across six categories (pedestrians, bicyclists, skateboarders, cars, carts and buses). In total, there are 60 video clips, for about

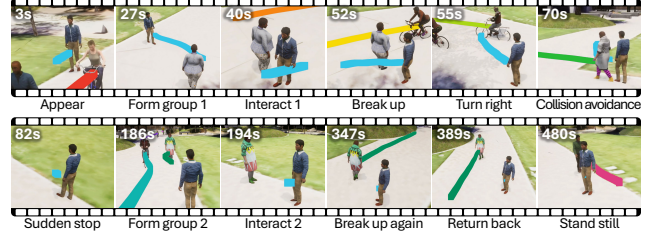


Figure 4. Visualization of the time-varying behavior changes (blue man). Our CrowdES can autonomously generate realistic, long-term behavioral sequences for each agent in a scene.

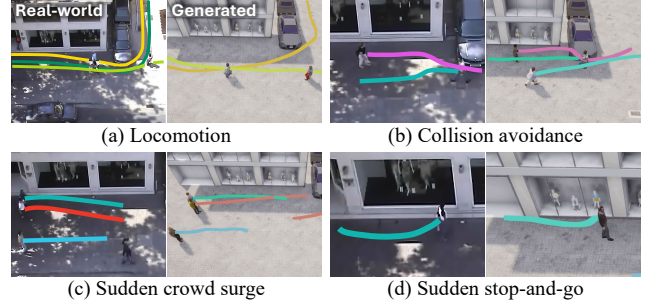


Figure 5. Visualization of the generated behaviors compared to the real-world behaviors in the ZARA1 scene.

5 hours, captured on a university campus. Following [66], 17 scenes are used for evaluation, while the remainder are used for training. For ETH-UCY and SDD, their test splits are the unseen environments over training sets.

The GCS dataset provides 1.11 hours of video taken in a highly congested terminal capturing 13,394 pedestrians. For evaluation, we divide the video into an 80%-20% train-test split. The EDIN dataset consists of 118 video clips, 873 hours and tracking 108,993 pedestrians, captured at another university campus. We use the clips recorded in December for evaluation, while the remainder are utilized for training.

Evaluation metrics. There have been various methodologies proposed to assess crowd behaviors, including entropy-based similarity metrics [49, 61], agent-level accuracy, realism, and certainty metrics [153], and group frequency and density metrics [113]. These approaches inherently require a shared observation state and manually annotated group labels, making them unsuitable for our framework. Meanwhile, Itatani *et al.* [53] suggest a user-study-based realism measure, which is impractical for evaluating our long videos. To address these limitations, we introduce eight evaluation metrics designed for our crowd behavior generation benchmark, incorporating key insights from previous studies.

In order to evaluate scene-level similarity, we define realism metrics based on the Earth Mover’s Distance (EMD) [139]. Inspired by the quadrat sampling method [124], we measure the distribution similarity of crowds with respect to the agent types. Specifically, we calculate density (Dens.), frequency (Freq.), and coverage (Cov.) every second on a 10×10 grid for both the gener-

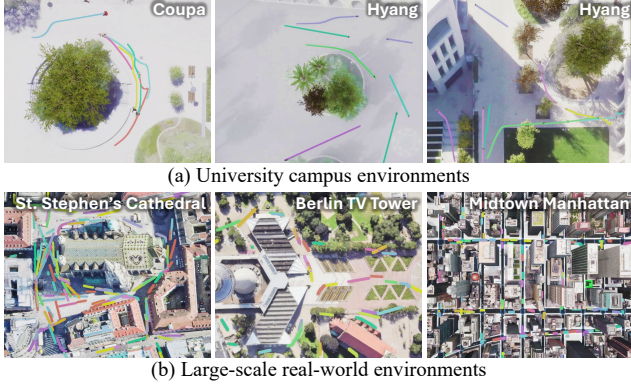


Figure 6. Visualization of the flexibility of our CrowdES framework across diverse university campuses and large-scale real-world environments (Line: 5-second future trajectories).

ated and ground-truth test sets, and then compute the EMD between them. Similarly, we assess population similarity (Pop.) between the generated and ground-truth histograms of agent counts, collected every second.

To estimate the agent-level similarity, we introduce a kinematics (Kinem.) metric, which averages four EMDs of travel velocities, accelerations, distances and times for all agents in the generated and ground-truth sets. To measure the spatial accuracy of trajectories at meter scale, we calculate the average minimum pairwise Dynamic Time Warping (DTW) distance [141] between the generated and ground-truth trajectories. Diversity (Div.) is evaluated by examining how comprehensively these minimum-distance pairs cover the full set of agent trajectories. Lastly, we use collision rate (Col.) to check the percentage of generated cases where agents collide with each other.

Evaluation methodology. Starting from scene images in the test datasets, the comparison models are intended to generate scenarios composed of multiple crowd trajectories. To evaluate the metrics, the generated scenarios are truncated to have the same duration as the ground-truth video. For a fair comparison, each experiment is repeated 20 times, and the results are averaged to reduce randomness and enhance numerical stability.

4.2. Evaluation Results

We compare our CrowdES framework with two baseline models, including algorithmic and learnable methods. SE-ORCA combines a random surface emitter [151] and ORCA [171] algorithms to implement the continuous crowd emerging and collision-avoidance actions. VAE [97] uses a conditional variational auto-encoder, in place of the conditional diffusion model in the crowd emitter, and SDS in the crowd simulator, to learn the distributions of crowd behavior dynamics.

In Tab. 1, we report the comparison results with respect to both scene-level realism and agent-level accuracy. The

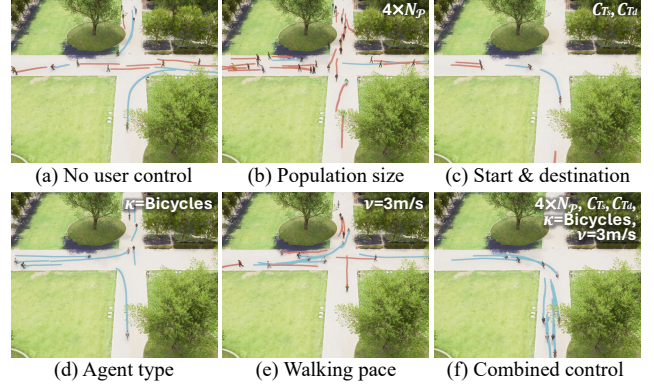


Figure 7. Visualization of the controllability of our CrowdES framework with various user controls (Red: Pedestrians, Blue: Bicycles).

results demonstrate that our CrowdES consistently outperforms these methods across nearly all metrics and datasets. In particular, our diffusion-based crowd emitter successfully populates the environments with a highly realistic distribution of agents while facilitating group formation. At the same time, the generated agents exhibit realistic kinematics and follow trajectories that closely resemble actual paths. Additionally, compared to the conventional algorithmic approach [171], the probabilistic behavior-switching mechanism significantly enhances the diversity of the generated paths. However, CrowdES occasionally encounters a few more collisions than SE-ORCA [171], focusing on collision avoidance ability. This is because our model concentrates on diverse behavioral switching motions. Nevertheless, since the collisions are reported as percentage units, our crowd simulator, which has neighborhood awareness, typically achieves safe locomotion trajectories. We note that CrowdES generates crowd behaviors in real-time, typically requiring 48 seconds to generate a one-hour scenario.

4.3. Flexibility and Controllability

We visualize CrowdES’s ability to generate diverse crowd behaviors by presenting our results alongside several real-world cases. To aid visualization, we synthesize virtual crowds using the CARLA simulator [34]. In Fig. 4, we track an individual throughout the generated long-term scenario to show the socially acceptable behavioral diversity. Our CrowdES framework produces realistic, long-term behaviors with diverse intermediate interactions using probabilistic state transitions within a Markov chain. In Fig. 5, our crowd simulator successfully produces natural, environment-aware paths from starting points to destinations, which is essential for crowd behavior modeling. Collision avoidance is another key element, and our network successfully learns this behavior in a data-driven manner. Notably, CrowdES is capable of replicating sudden crowd behaviors, such as sudden surges and stop-and-go motions, without any supervision.

Next, we demonstrate the flexibility of our framework

Model	Component	Dens.	Freq.	Cov.	Pop.
Crowd Emitter	w/o Diffusion model	0.052	0.052	0.047	0.678
	w/o Population sampling	0.051	0.042	0.040	0.565
	w/o Spatial layout condition	<u>0.039</u>	0.039	0.037	0.486
	w/o Collectivity transformer	<u>0.039</u>	<u>0.037</u>	<u>0.033</u>	<u>0.478</u>
	CrowdES	0.038	0.033	0.030	0.463

Table 2. Ablation study of the crowd emitter.

Model	Component	Kinem.	DTW	Div.	Col.
Crowd Simulator	w/o Spatial layout condition	0.671	6.703	<u>0.350</u>	1.554
	w/o Navigation mesh	0.619	6.658	0.348	<u>0.602</u>
	w/o Social interaction	0.709	6.937	0.334	2.709
	w/o State-switching	0.706	6.351	0.343	1.714
	CrowdES	<u>0.650</u>	<u>6.352</u>	0.354	0.411

Table 3. Ablation study of the crowd simulator.

to populate crowds and to make their movements within real-world environments in Fig. 6(a). In a university campus scene, virtual crowds naturally emerge from buildings or scene boundaries and move toward their destinations. During their locomotion toward destinations, crowds avoid collisions with both environmental obstacles like trees and bushes, and other agents. To emphasize the robustness of our CrowdES, we design three more large-scale and complex real-world environments in Fig. 6(b). In the Cathedral scene, our crowd simulator generates smooth locomotion paths between buildings. In the TV tower scene, our framework produces realistic crowd behaviors within the intricate structures. Lastly, in the extremely challenging Manhattan scenario, we observe that agents successfully navigate and populate the dense, large-scale urban landscape.

Lastly, we demonstrate the controllability of our CrowdES framework in Fig. 7. Starting from an initial uncontrolled scenario, users can adjust scene-level parameters, including the overall size of population and start/destination areas. At the agent level, users can also edit all agent parameters, for example, changing the agent types from pedestrians to bicycles or accelerating each agent’s pace. These components are customizable and can be adjusted simultaneously. As a result, CrowdES not only fully automatically synthesizes crowd behaviors in complex environments, but also offers controllability for various applications.

4.4. Ablation Studies

Component of the crowd emitter. We conduct an ablation study for the crowd emitter on the SDD dataset by removing its components one-by-one, in Tab. 2. First, we evaluate the backbone generative model. We determine that incorporating the powerful generative capabilities of diffusion models produces a more realistic population distribution compared to VAE models. Next, we replace the population probability prediction with a population regression network. The regression approach fails to capture dynamic population shifts, such as sudden crowd surges, which decreases the scene-level similarities. Third, we remove the spatial layout, conditioned by the diffusion models. We confirm that the

# Behavior States	Scene-Level Realism				Agent-Level Accuracy			
	Dens.	Freq.	Cov.	Pop.	Kinem.	DTW	Div.	Col.
$B = 1$	0.041	0.038	0.035	0.513	0.706	6.351	0.343	1.735
$B = 2$	<u>0.039</u>	<u>0.035</u>	0.032	0.486	0.667	6.256	0.352	1.030
$B = 4$	0.038	0.033	<u>0.031</u>	<u>0.464</u>	<u>0.656</u>	<u>6.236</u>	0.349	0.825
$B = 8$	0.038	0.033	0.030	0.463	0.650	6.352	<u>0.354</u>	<u>0.411</u>
$B = 16$	0.038	0.033	<u>0.031</u>	<u>0.464</u>	<u>0.656</u>	<u>6.236</u>	0.353	0.333
$B = 32$	<u>0.039</u>	<u>0.035</u>	0.032	0.476	0.674	6.225	0.357	0.497

Table 4. Ablation study on the number of behavior states.

appearance and population density maps, commonly used as user controls in conventional software, are also beneficial within a learnable framework. Lastly, we examine the independent agent parameter generation in place of the collective generation. In particular, our collective generation approach improves the performance of the frequency and coverage metrics, which are sensitive to group cohesion.

Components of the crowd simulator. Next, we conduct another ablation study of the crowd simulator in Tab. 3. First, the spatial layout and navigation mesh effectively constrain agents to traversable areas, guiding them to plan paths more closely aligned with real trajectories according to the DTW metric. We then remove the social interaction module, which leads to a reduction in diversity and increases collision cases. We also evaluate the state-switching module. The dynamic behavior transitions enable us to simulate unpredictable behaviors, such as sudden stops and stop-and-go motions, which improve both kinematic similarity and behavioral diversity. Lastly, we explore the impact of the number of behavior states for optimal performance. As shown in Tab. 4, $B = 8$ states achieve the best performance. While additional behavior states may enhance diversity and improve the likelihood of corresponding real trajectories, excessive behavior transitions lead to lower kinematic fidelity.

5. Conclusion

In this paper, we present CrowdES, a framework for generating continuous and realistic crowd trajectories with diverse behaviors from single input images. By combining the crowd emitter which assigns individual attributes with the crowd simulator that produces detailed trajectories, our method captures complex interactions and heterogeneity among crowds. Our framework is also user-controllable, allowing customization of parameters such as population density and walking speed. In addition, we introduce a new evaluation protocol for continuous crowd generation tasks. Through a variety of experiments, we demonstrate that our CrowdES generates lifelike crowd behaviors with respect to both scene-level realism and individual trajectory accuracy across diverse environments for dynamic crowd simulation.

Acknowledgement This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(RS-2024-00338439), Institute of Information & communications Technology Planning & Evaluation (IITP) grant (RS-2021-II212068, Artificial Intelligence Innovation Hub), and the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land Infrastructure and Transport (Grant RS-2023-00256888).

References

- [1] Guy Ackerson and K Fu. On state estimation in switching environments. *IEEE transactions on automatic control*, 1970. 3
- [2] Alexandre Alahi, Kratharth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2
- [3] Abdul Fatir Ansari, Konstantinos Benidis, Richard Kurle, Ali Caner Turkmen, Harold Soh, Alexander J Smola, Bernie Wang, and Tim Januschowski. Deep explicit duration switching models for time series. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2021. 3
- [4] Gökay Aydemir, Adil Kaan Akan, and Fatma Güney. Adapt: Efficient multi-agent trajectory prediction with adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [5] Inhwon Bae and Hae-Gon Jeon. Disentangled multi-relational graph convolutional network for pedestrian trajectory prediction. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2
- [6] Inhwon Bae and Hae-Gon Jeon. A set of control points conditioned pedestrian trajectory prediction. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2
- [7] Inhwon Bae, Jin-Hwi Park, and Hae-Gon Jeon. Learning pedestrian group representations for multi-modal trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2
- [8] Inhwon Bae, Jin-Hwi Park, and Hae-Gon Jeon. Non-probability sampling network for stochastic human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [9] Inhwon Bae, Jean Oh, and Hae-Gon Jeon. EigenTrajectory: Low-rank descriptors for multi-modal trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 5
- [10] Inhwon Bae, Junoh Lee, and Hae-Gon Jeon. Can language beat numerical regression? language-based multimodal trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [11] Inhwon Bae, Young-Jae Park, and Hae-Gon Jeon. Singulartrajectory: Universal trajectory predictor using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [12] Carles Balsells-Rodas, Yixin Wang, Pedro AM Mediano, and Yingzhen Li. Identifying nonstationary causal structures with high-order markov switching models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 3
- [13] Yaakov Bar-Shalom and Xiao-Rong Li. Estimation and tracking: Principles, techniques, and software. *IEEE Antennas and Propagation Magazine*, 1996. 3
- [14] Apratim Bhattacharyya, Michael Hanselmann, Mario Fritz, Bernt Schiele, and Christoph-Nikolas Straehle. Conditional flow variational autoencoders for structured sequence prediction. *arXiv preprint arXiv:1908.09008*, 2020. 2
- [15] Niccoló Bisagno, Bo Zhang, and Nicola Conci. Group lstm: Group trajectory prediction in crowded scenarios. In *Proceedings of the European Conference on Computer Vision Workshop (ECCVW)*, 2018. 2
- [16] Eric Bouvier, Eyal Cohen, and Laurent Najman. From crowd simulation to airbag deployment: particle systems, a new paradigm of simulation. *Journal of Electronic imaging*, 1997. 2
- [17] Sergio Casas, Ben Agro, Jiageng Mao, Thomas Gilles, Alexander Cui, Thomas Li, and Raquel Urtasun. Detra: A unified model for object detection and trajectory forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [18] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2019. 5
- [19] Panayiotis Charalambous, Julien Pettre, Vassilis Vassiliades, Yiorgos Chrysanthou, and Nuria Pelechano. Greil-crowds: crowd simulation with deep reinforcement learning and examples. *ACM Transactions on Graphics (TOG)*, 2023. 1, 3
- [20] Guangyi Chen, Junlong Li, Jiwen Lu, and Jie Zhou. Human trajectory prediction via counterfactual analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [21] Guangyi Chen, Junlong Li, Nuoxing Zhou, Liangliang Ren, and Jiwen Lu. Personalized trajectory prediction via distribution discrimination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [22] Guangyi Chen, Zhenhao Chen, Shunxing Fan, and Kun Zhang. Unsupervised sampling promoting for stochastic human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [23] Hao Chen, Jiaze Wang, Kun Shao, Furui Liu, Jianye Hao, Chenyong Guan, Guangyong Chen, and Pheng-Ann Heng. Traj-mae: Masked autoencoders for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [24] Hongyi Chen, Jingtao Ding, Yong Li, Yue Wang, and Xiaoping Zhang. Social physics informed diffusion model for crowd simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024. 3
- [25] Sehwan Choi, Jungho Kim, Junyong Yun, and Jun Won Choi. R-pred: Two-stage motion prediction via tube-query attention-based trajectory refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [26] Sourav Das, Guglielmo Camporese, and Lamberto Ballan. Distilling knowledge for short-to-long term trajectory predic-

- tion. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024. 2
- [27] Patrick Dendorfer, Aljosa Osep, and Laura Leal-Taixe. Goalgan: Multimodal trajectory prediction based on goal position estimation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 2
- [28] Patrick Dendorfer, Sven Elflein, and Laura Leal-Taixé. Mgan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [29] Nachiket Deo and Mohan M. Trivedi. Trajectory forecasts in unknown environments conditioned on grid-based plans. *arXiv preprint arXiv:2001.00735*, 2020. 2
- [30] Jeevan Devaranjan, Amlan Kar, and Sanja Fidler. Meta-sim2: Unsupervised learning of scene structure for synthetic data generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3
- [31] Wenhao Ding, Yulong Cao, Ding Zhao, Chaowei Xiao, and Marco Pavone. Realgen: Retrieval augmented generation for controllable traffic scenarios. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1, 3
- [32] Yonghao Dong, Le Wang, Sanping Zhou, and Gang Hua. Sparse instance conditioned multimodal trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [33] Zhe Dong, Bryan Seybold, Kevin Murphy, and Hung Bui. Collapsed amortized variational inference for switching nonlinear dynamical systems. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 3
- [34] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, 2017. 7
- [35] Funda Durupinar, Nuria Pelechano, Jan Allbeck, Uğur Güdükbay, and Norman I Badler. How the ocean personality model affects the perception of crowds. *IEEE Computer Graphics and Applications*, 2009. 3
- [36] Epic Games Inc. Niagara, Unreal Engine, Version 5.5. <https://www.unrealengine.com>, 2024. 1, 3
- [37] Yixuan Fan, Yali Li, and Shengjin Wang. Risk-aware self-consistent imitation learning for trajectory planning in autonomous driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [38] Lan Feng, Quanyi Li, Zhenghao Peng, Shuhan Tan, and Bolei Zhou. Trafficgen: Learning to generate diverse and realistic traffic scenarios. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1, 3
- [39] Lan Feng, Mohammadhossein Bahari, Kaouther Mes-saoud Ben Amor, Éloi Zablocki, Matthieu Cord, and Alexandre Alahi. Unitraj: A unified framework for scalable vehicle trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [40] Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *Neural Networks*, 108:466–478, 2018. 2
- [41] Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2017. 3
- [42] Ryo Fujii, Hideo Saito, and Ryo Hachiuma. Realtraj: Towards real-world pedestrian trajectory forecasting. *arXiv preprint arXiv:2411.17376*, 2024. 2
- [43] Yang Gao, Po-Chien Luan, and Alexandre Alahi. Multi-transmotion: Pre-trained model for human motion prediction. *Proceedings of the Conference on Robot Learning (CoRL)*, 2024. 2
- [44] Zoubin Ghahramani and Geoffrey E Hinton. Variational learning for switching state-space models. *Neural computation*, 2000. 3
- [45] Golaem. Golaem for MAYA, Version 9.1.1. <https://golaem.com>, 2024. 3
- [46] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [47] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6
- [48] Stephen J Guy, Sujeong Kim, Ming C Lin, and Dinesh Manocha. Simulating heterogeneous crowd behaviors using personality trait theory. In *Proceedings of ACM SIGGRAPH*, 2011. 3
- [49] Stephen J Guy, Jur Van Den Berg, Wenxi Liu, Rynson Lau, Ming C Lin, and Dinesh Manocha. A statistical similarity measure for aggregate crowd dynamics. *ACM Transactions on Graphics (TOG)*, 2012. 6
- [50] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 1, 2
- [51] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2
- [52] Ronny Hug, Wolfgang Hübner, and Michael Arens. Introducing probabilistic bézier curves for n-step sequence prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2
- [53] Reiya Itatani and Nuria Pelechano. Social crowd simulation: Improving realism with social rules and gaze behavior. In *Proceedings of ACM SIGGRAPH*, 2024. 6
- [54] Itrch, Shekn. Path Finder. <https://github.com/Tugcga/Path-Finder>, 2021. 5
- [55] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

- [56] Jaewoo Jeong, Daehee Park, and Kuk-Jin Yoon. Multi-agent long-term 3d human pose forecasting via interaction-aware trajectory conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [57] Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [58] Ruochen Jiao, Xiangguo Liu, Takami Sato, Qi Alfred Chen, and Qi Zhu. Semi-supervised semantics-guided adversarial training for robust trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [59] Eunjung Ju, Myung Geol Choi, Minji Park, Jehee Lee, Kang Hoon Lee, and Shigeo Takahashi. Morphable crowds. *ACM Transactions on Graphics (TOG)*, 2010. 3
- [60] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3
- [61] Ioannis Karamouzas, Nick Sohre, Ran Hu, and Stephen J Guy. Crowd space: a predictive crowd analysis technique. *ACM Transactions on Graphics (TOG)*, 2018. 6
- [62] Sujeong Kim, Stephen J Guy, Dinesh Manocha, and Ming C Lin. Interactive simulation of dynamic crowd behaviors using general adaptation syndrome theory. In *Proceedings of ACM SIGGRAPH*, 2012. 3
- [63] Sungjune Kim, Hyung-gun Chi, Hyerin Lim, Karthik Raman, Jinkyu Kim, and Sangpil Kim. Higher-order relational reasoning for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [64] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 2
- [65] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Socialbigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2019. 2
- [66] Parth Kothari, Sven Kreiss, and Alexandre Alahi. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, 2021. 6
- [67] Parth Kothari, Brian Siffringer, and Alexandre Alahi. Interpretable social anchors for human trajectory forecasting in crowds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5
- [68] Richard Kurl, Syama Sundar Rangapuram, Emmanuel de Bézenac, Stephan Günnemann, and Jan Gasthaus. Deep rao-blackwellised particle filters for time series forecasting. *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2020. 3
- [69] Taesoo Kwon, Kang Hoon Lee, Jehee Lee, and Shigeo Takahashi. Group motion editing. In *Proceedings of ACM SIGGRAPH*, 2008. 3
- [70] Yu-Chi Lai, Stephen Chenney, and ShaoHua Fan. Group motion graphs. In *Proceedings of ACM SIGGRAPH*, 2005. 3
- [71] Jaedong Lee, Jungdam Won, and Jehee Lee. Crowd simulation by deep reinforcement learning. In *Proceedings of ACM SIGGRAPH*, 2018. 1, 3
- [72] Kang Hoon Lee, Myung Geol Choi, and Jehee Lee. Motion patches: building blocks for virtual environments annotated with motion data. In *Proceedings of ACM SIGGRAPH*, 2006. 1
- [73] Kang Hoon Lee, Myung Geol Choi, Qyoun Hong, and Jehee Lee. Group behavior from video: a data-driven approach to crowd simulation. In *Proceedings of ACM SIGGRAPH*, 2007. 3
- [74] Mihee Lee, Samuel S. Sohn, Seonghyeon Moon, Sejong Yoon, Mubbasir Kapadia, and Vladimir Pavlovic. Muse-vae: Multi-scale vae for environment-aware long term trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [75] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2
- [76] Seongju Lee, Junseok Lee, Yeonguk Yu, Taeri Kim, and Kyobin Lee. Mart: Multiscale relational transformer networks for multi-agent trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [77] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. *Computer Graphics Forum*, 26(3): 655–664, 2007. 3, 6
- [78] Alon Lerner, Yiorgos Chrysanthou, Ariel Shamir, and Daniel Cohen-Or. Context-dependent crowd evaluation. In *Computer Graphics Forum*, 2010. 3
- [79] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Conditional generative neural system for probabilistic trajectory prediction. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019. 2
- [80] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2020. 2
- [81] Shijie Li, Yanying Zhou, Jinhui Yi, and Juergen Gall. Spatial-temporal consistency network for low-latency trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [82] Yuke Li. Which way are you going? imitative decision learning for path forecasting in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [83] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

- [84] Junwei Liang, Lu Jiang, and Alexander Hauptmann. Simaug: Learning robust representations from simulation for trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2
- [85] Junwei Liang, Lu Jiang, Kevin Murphy, Ting Yu, and Alexander Hauptmann. The garden of forking paths: Towards multi-future trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [86] Rongqin Liang, Yuanman Li, Xia Li, Yi Tang, Jiantao Zhou, and Wenbin Zou. Temporal pyramid network for pedestrian trajectory prediction with multi-supervision. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2
- [87] Daniel Liberzon. *Switching in systems and control*. Springer, 2003. 3
- [88] Xiaotong Lin, Tianming Liang, Jianhuang Lai, and Jianfang Hu. Progressive pretext task learning for human trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [89] Scott W Linderman, Andrew C Miller, Ryan P Adams, David M Blei, Liam Paninski, and Matthew J Johnson. Recurrent switching linear dynamical systems. *arXiv preprint arXiv:1610.08466*, 2016. 3
- [90] Yuejiang Liu, Qi Yan, and Alexandre Alahi. Social nce: Contrastive learning of socially-aware motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [91] Yongtuo Liu, Sara Magliacane, Miltiadis Kofinas, and Efstratios Gavves. Graph switching dynamical systems. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. 3
- [92] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 5
- [93] Jack Lu, Kelvin Wong, Chris Zhang, Simon Suo, and Raquel Urtasun. Scenecontrol: Diffusion for controllable traffic scene generation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 1, 3
- [94] Yuexin Ma, Xinge Zhu, Xinjing Cheng, Ruigang Yang, Jiming Liu, and Dinesh Manocha. Autotrajjectory: Label-free trajectory extraction and prediction from videos using dynamic points. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2
- [95] Takahiro Maeda and Norimichi Ukita. Fast inference and update of probabilistic density estimation on trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [96] Barbara Majecka. Statistical models of pedestrian behaviour in the forum. Master's Thesis, School of Informatics, University of Edinburgh, 2009. 6
- [97] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 7
- [98] Karttikeya Mangalam, Yang An, Harshayu Girase, and Jitendra Malik. From goals, waypoints & paths to long term human trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [99] Huynh Manh and Gita Alaghband. Scene-1stm: A model for human trajectory prediction. *arXiv preprint arXiv:1808.04018*, 2018.
- [100] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [101] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Mantra: Memory augmented networks for multiple trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [102] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Multiple trajectory prediction of moving agents with memory augmented networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 2
- [103] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Smemo: Social memory for trajectory forecasting. *arXiv preprint arXiv:2203.12446*, 2022. 2
- [104] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2
- [105] Abdullah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [106] Abdullah Mohamed, Deyao Zhu, Warren Vu, Mohamed Elhoseiny, and Christian Claudel. Social-implicit: Rethinking trajectory prediction evaluation and the effectiveness of implicit maximum likelihood estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2
- [107] Mononen, Mikko. Recast Navigation. <https://github.com/memononen/recastnavigation>, 2014. 5
- [108] Alessio Monti, Angelo Porrello, Simone Calderara, Pasquale Coscia, Lamberto Ballan, and Rita Cucchiara. How many observations are enough? knowledge distillation for trajectory forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [109] Seokha Moon, Hyun Woo, Hongbeen Park, Haeji Jung, Reza Mahjourian, Hyung-gun Chi, Hyerin Lim, Sangpil Kim, and Jinkyu Kim. Visiontrap: Vision-augmented trajectory prediction guided by textual descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [110] Mehdi Moussaïd, Niriasca Perozo, Simon Garnier, Dirk Helbing, and Guy Theraulaz. The walking behaviour of

- pedestrian social groups and its impact on crowd dynamics. *Public Library of Science One*, 2010. 4
- [111] Ingrid Navarro and Jean Oh. Social-patternn: Socially-aware trajectory prediction guided by motion patterns. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022. 2
- [112] Sang Min Oh, Ananth Ranganathan, James M Rehg, and Frank Dellaert. A variational inference method for switching linear dynamic systems. *Georgia Institute of Technology*, 2005. 3
- [113] Stuart O'Connor, James Shuttleworth, Simon Colreavy-Donnelly, and Fotis Liarokapis. Assessing the perceived realism of agent grouping dynamics for adaptation and simulation. *Entertainment Computing*, 2019. 6
- [114] Andreas Panayiotou, Theodoros Kyriakou, Marilena Lemonari, Yiorgos Chrysanthou, and Panayiotis Charalambous. Ccp: Configurable crowd profiles. In *Proceedings of ACM SIGGRAPH*, 2022. 1, 3
- [115] Bo Pang, Tianyang Zhao, Xu Xie, and Ying Nian Wu. Trajectory prediction with latent belief energy-based model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [116] Daehee Park, Hobin Ryu, Yunseo Yang, Jegyeong Cho, Jiwon Kim, and Kuk-Jin Yoon. Leveraging future relationship reasoning for vehicle trajectory prediction. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2
- [117] Daehee Park, Jaewoo Jeong, and Kuk-Jin Yoon. Improving transferability for cross-domain trajectory prediction via neural stochastic differential equation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024. 2
- [118] Daehee Park, Jaeseok Jeong, Sung-Hoon Yoon, Jaewoo Jeong, and Kuk-Jin Yoon. T4p: Test-time training of trajectory prediction via masked autoencoder and actor-specific token memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [119] Young-Jae Park, Minseok Seo, Doyi Kim, Hyeri Kim, Sanghoon Choi, Beomkyu Choi, Jeongwon Ryu, Sohee Son, Hae-Gon Jeon, and Yeji Choi. Long-term typhoon trajectory prediction: A physics-conditioned approach without reanalysis data. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2
- [120] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2009. 2, 6
- [121] Julien Pettr . Populate your game scene. In *International Workshop on Motion in Games*, 2008. 3
- [122] Mark Pfeiffer, Giuseppe Paolo, Hannes Sommer, Juan I. Nieto, Roland Y. Siegwart, and C sar Cadena. A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 2
- [123] Jonah Philion, Xue Bin Peng, and Sanja Fidler. Trajeglish: Traffic modeling as next-token prediction. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 2
- [124] Roscoe Pound and Frederic E Clements. Ii. a method of determining the abundance of secondary species. *Minnesota Botanical Studies*, 1898. 6
- [125] Mozghan Pourkeshavarz, Changhe Chen, and Amir Rasouli. Learn tarot with mentor: A meta-learned self-supervised approach for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [126] Mozghan Pourkeshavarz, Junrui Zhang, and Amir Rasouli. Cadet: a causal disentanglement approach for robust trajectory prediction in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [127] Mozghan Pourkeshavarz, Junrui Zhang, and Amir Rasouli. Dyset: A dynamic masked self-distillation approach for robust trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [128] Ethan Pronovost, Meghana Reddy Ganesina, Noureldin Hendy, Zeyu Wang, Andres Morales, Kai Wang, and Nick Roy. Scenario diffusion: Controllable driving scenario generation with diffusion. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2023. 1, 3
- [129] Ethan Pronovost, Kai Wang, and Nick Roy. Generating driving scenes with diffusion. In *Proceedings of the IEEE International Conference on Robotics and Automation Workshop (ICRAW)*, 2023. 3
- [130] Eike Rehder and Horst Kloeden. Goal-directed pedestrian prediction. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015. 2
- [131] Eike Rehder, Florian Wirth, Martin Lauer, and Christoph Stiller. Pedestrian prediction by planning using deep neural networks. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 2
- [132] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3
- [133] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4
- [134] Zhiguo Ren, Panayiotis Charalambous, Julien Bruneau, Qunsheng Peng, and Julien Pettr . Group modeling: A unified velocity-based approach. In *Computer Graphics Forum*, 2017. 3
- [135] Craig W Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 1987. 1, 3
- [136] Craig W Reynolds et al. Steering behaviors for autonomous characters. In *Game developers conference*, 1999. 3
- [137] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human

- trajectory understanding in crowded scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 6
- [138] Sam Roweis and Zoubin Ghahramani. A unifying review of linear gaussian models. *Neural computation*, 1999. 3
- [139] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1998. 6
- [140] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [141] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 2007. 7
- [142] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2
- [143] Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [144] Nasim Shafiee, Taskin Padir, and Ehsan Elhamifar. Introvert: Human trajectory prediction via conditional 3d attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [145] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [146] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Fang Zheng, Nanning Zheng, and Gang Hua. Social interpretable tree for pedestrian trajectory prediction. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 2
- [147] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Wei Tang, Nanning Zheng, and Gang Hua. Representing multimodal behaviors with mean location for pedestrian trajectory prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023. 2
- [148] Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [149] Xiaodan Shi, Xiaowei Shao, Zipei Fan, Renhe Jiang, Haoran Zhang, Zhiling Guo, Guangming Wu, Wei Yuan, and Ryosuke Shibasaki. Multimodal interaction-aware trajectory prediction in crowded space. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2
- [150] Xiaodan Shi, Xiaowei Shao, Guangming Wu, Haoran Zhang, Zhiling Guo, Renhe Jiang, and Ryosuke Shibasaki. Social-dpf: Socially acceptable distribution prediction of futures. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2
- [151] Side Effects Software Inc. Houdini, Version 20.5. <https://www.sidefx.com>, 2024. 3, 7
- [152] Greg Snook. Simplified 3d movement and pathfinding using navigation meshes. *Game programming gems*, 2000. 5
- [153] Samuel S Sohn, Mihee Lee, Seonghyeon Moon, Gang Qiao, Muhammad Usman, Sejong Yoon, Vladimir Pavlovic, and Mubbasir Kapadia. A2x: An end-to-end framework for assessing agent and environment interactions in multimodal human trajectory prediction. *Computers & Graphics*, 2022. 6
- [154] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5
- [155] Hao Sun, Zhiqun Zhao, and Zhihai He. Reciprocal learning networks for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [156] Jianhua Sun, Qinhong Jiang, and Cewu Lu. Recursive social behavior graph for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2
- [157] Jianhua Sun, Yuxuan Li, Hao-Shu Fang, and Cewu Lu. Three steps to multimodal trajectory prediction: Modality clustering, classification and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [158] Jianhua Sun, Yuxuan Li, Liang Chai, Hao-Shu Fang, Yong-Lu Li, and Cewu Lu. Human trajectory prediction with momentary observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [159] Jianhua Sun, Yuxuan Li, Liang Chai, and Cewu Lu. Stimulus verification is a universal and effective sampler in multimodal human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [160] Shuo Sun, Zekai Gu, Tianchen Sun, Jiawei Sun, Chengran Yuan, Yuhang Han, Dongen Li, and Marcelo H Ang. Drivescenegen: Generating diverse and realistic driving scenarios from scratch. In *IEEE Robotics and Automation Letters (RA-L)*, 2024. 1, 3
- [161] Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Scenegen: Learning to generate realistic traffic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 3
- [162] Shuhan Tan, Boris Ivanovic, Xinshuo Weng, Marco Pavone, and Philipp Kraehenbuehl. Language conditioned traffic generation. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023. 3
- [163] Xiaolong Tang, Meina Kan, Shiguang Shan, Zhilong Ji, Jinfeng Bai, and Xilin Chen. Hpnet: Dynamic trajectory

- forecasting with historical prediction attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [164] Chaofan Tao, Qinhong Jiang, and Lixin Duan. Dynamic and static context-aware lstm for multi-agent motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2
- [165] Neerja Thakkar, Karttikeya Mangalam, Andrea Bajcsy, and Jitendra Malik. Adaptive human trajectory prediction via latent corridors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [166] Paul Tozour and IS Austin. Building a near-optimal navigation mesh. *AI game programming wisdom*, 2002. 5
- [167] Li-Wu Tsao, Yan-Kai Wang, Hao-Siang Lin, Hong-Han Shuai, Lai-Kuan Wong, and Wen-Huang Cheng. Social-ssl: Self-supervised cross-sequence representation learning based on transformers for multi-agent trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2
- [168] uCrowds. SimCrowds 2025, Version 1.6. <https://www.ucrowds.com>, 2024. 1, 3
- [169] Branislav Ulicny, Pablo de Heras Ciechomski, and Daniel Thalmann. Crowdbush: interactive authoring of real-time crowd scenes. In *Proceedings of ACM SIGGRAPH*, 2004. 1
- [170] Jur Van den Berg, Ming Lin, and Dinesh Manocha. Reciprocal velocity obstacles for real-time multi-agent navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2008. 1, 3
- [171] Jur Van Den Berg, Stephen J Guy, Jamie Snape, Ming Lin, and Dinesh Manocha. Optimal reciprocal collision avoidance for multi-agent navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation Workshop (ICRAW)*, 2010. 3, 7
- [172] Wouter Van Toll, Roy Triesscheijn, Marcelo Kallmann, Ramon Oliva, Nuria Pelechano, Julien Pettr , and Roland Ger-aerts. A comparative study of navigation meshes. In *Proceedings of the 9th International Conference on Motion in Games*, 2016. 5
- [173] Daksh Varshneya and G. Srinivasaraghavan. Human trajectory prediction using spatially aware deep attention models. *arXiv preprint arXiv:1705.09436*, 2017. 2
- [174] Petar Veli kovi , Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Li , and Yoshua Bengio. Graph attention networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. 2
- [175] Anirudh Vemula, Katharina Muelling, and Jean Oh. Social attention: Modeling attention in human crowds. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 2
- [176] Chuhua Wang, Yuchen Wang, Mingze Xu, and David J Crandall. Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters (RA-L)*, 2022. 2
- [177] Jihui-Ciang Wu Wang, Hong-Han Shuai, and Wen-Huang Cheng. Trajprompt: Aligning color trajectory with vision-language representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [178] Yuning Wang, Pu Zhang, Lei Bai, and Jianru Xue. Fend: A future enhanced distribution-aware contrastive learning framework for long-tail trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [179] Yixiao Wang, Chen Tang, Lingfeng Sun, Simone Rossi, Yichen Xie, Chensheng Peng, Thomas Hannagan, Stefano Sabatini, Nicola Poerio, Masayoshi Tomizuka, et al. Optimizing diffusion models for joint trajectory prediction and controllable generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [180] Di Wen, Haoran Xu, Zhaocheng He, Zhe Wu, Guang Tan, and Peixi Peng. Density-adaptive model based on motif matrix for multi-agent trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [181] Song Wen, Hao Wang, and Dimitris Metaxas. Social ode: Multi-agent trajectory forecasting with neural ordinary differential equations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2
- [182] Conghao Wong, Beihao Xia, Ziming Hong, Qinmu Peng, Wei Yuan, Qiong Cao, Yibo Yang, and Xinge You. View vertically: A hierarchical network for trajectory prediction via fourier spectrums. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.
- [183] Conghao Wong, Beihao Xia, Ziqian Zou, Yulong Wang, and Xinge You. Socialcircle: Learning the angle-based social interaction representation for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [184] Enze Xie, Wenhao Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2021. 4
- [185] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [186] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [187] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. Eqmotion: Equivariant multi-agent motion prediction with invariant interaction reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [188] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2
- [189] Yi Xu and Yun Fu. Adapting to length shift: Flexilength network for trajectory prediction. In *Proceedings of the*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [190] Yi Xu, Jing Yang, and Shaoyi Du. Cf-lstm: Cascaded feature-based long short-term networks for predicting pedestrian trajectory. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2
- [191] Yi Xu, Lichen Wang, Yizhou Wang, and Yun Fu. Adaptive trajectory prediction via transferable gnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [192] Yi Xu, Armin Bazarjani, Hyung-gun Chi, Chiho Choi, and Yun Fu. Uncovering the missing pattern: Unified framework towards trajectory imputation and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [193] Hao Xue, Du Q Huynh, and Mark Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2018. 2
- [194] Kota Yamaguchi, Alexander C Berg, Luis E Ortiz, and Tamara L Berg. Who are you with and where are you going? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 2
- [195] Zhijie Yan, Pengfei Li, Zheng Fu, Shaocong Xu, Yongliang Shi, Xiaoxue Chen, Yuhang Zheng, Yang Li, Tianyu Liu, Chuxuan Li, et al. Int2: Interactive trajectory prediction at intersections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [196] Shanwen Yang, Tianrui Li, Xun Gong, Bo Peng, and Jie Hu. A review on crowd simulation and modeling. *Graphical Models*, 2020. 1
- [197] Yu Yao, Ella Atkins, Matthew Johnson-Roberson, Ram Vasudevan, and Xiaoxiao Du. Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. *IEEE Robotics and Automation Letters (RA-L)*, 2021. 2
- [198] Zhenzhen Yao, Guijuan Zhang, Dianjie Lu, and Hong Liu. Learning crowd behavior from real data: A residual network method for crowd simulation. *Neurocomputing*, 2020. 3
- [199] Barbara Yersin, Jonathan Maïm, Julien Pettr , and Daniel Thalmann. Crowd patches: populating large-scale virtual environments for real-time applications. In *Proceedings of the 2009 symposium on Interactive 3D graphics and games*, 2009. 1
- [200] Shuai Yi, Hongsheng Li, and Xiaogang Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6
- [201] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2
- [202] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [203] Jiangbei Yue, Dinesh Manocha, and He Wang. Human trajectory prediction via neural social physics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2
- [204] Haichao Zhang, Yi Xu, Hongsheng Lu, Takayuki Shimizu, and Yun Fu. Oostraj: Out-of-sight trajectory prediction with vision-positioning denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [205] Liang Zhang, Nathaniel Xu, Pengfei Yang, Gaojie Jin, Cheng-Chao Huang, and Lijun Zhang. Trajpac: Towards robustness verification of pedestrian trajectory prediction models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- [206] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [207] Pu Zhang, Jianru Xue, Pengfei Zhang, Nanning Zheng, and Wanli Ouyang. Social-aware pedestrian trajectory prediction via states refinement lstm. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.
- [208] He Zhao and Richard P. Wildes. Where are you heading? dynamic trajectory prediction with expert goal examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [209] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Benjamin Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, Congcong Li, and Dragomir Anguelov. Tnt: Target-driven trajectory prediction. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2020. 2
- [210] Mingbi Zhao, Wentong Cai, and Stephen John Turner. Clust: simulating realistic crowd behaviour by mining pattern from crowd videos. In *Computer graphics forum*, 2018. 3
- [211] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. Multi-agent tensor fusion for contextual trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [212] Bolei Zhou, Xiaogang Wang, and Xiaoou Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 6