IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Diversity-aware Crowd Model for Robust Robot Navigation in Human Populated Environment

Jiaxu Wu[1], Yusheng Wang[1], Tong Chen[1], Jun Jiang[2], Yongdong Wang[1], Qi An[1], Atsushi Yamashita[1]

***Abstract*—Robot navigation in human-populated environments poses challenges due to the diversity of human behaviors and the unpredictability of human paths. However, existing Reinforcement Learning (RL)-based methods often rely on simulators that lack sufficient diversity in human behavior, resulting in navigation policies that overfit specific human behavior and perform poorly in unseen environments. To address this, we propose a diversity-aware crowd model based on RL, employing Constrained Variational Exploration (VE) with a Mutual Information (MI)-based auxiliary reward to capture fine-grained behavioral diversity. The proposed model leverages a Centralized Training Decentralized Execution (CTDE) paradigm, which ensures stable exploration under multi-agent settings. Using the proposed diversity-aware model for training, we obtain robust robot navigation policies capable of handling diverse unseen scenarios. Extensive simulation and real-world experiments demonstrate the superior performance of our approach in achieving diverse crowd behaviors and enhancing robot navigation robustness. These findings highlight the potential of our method to advance safe and efficient robot operations in complex dynamic environments. For more details, please visit our project homepage https://wyd0817.github.io/project-diversity-awa/.**

***Index Terms*—Autonomous Vehicle Navigation, Human-Aware Motion Planning, Reinforcement Learning, Collision Avoidance, Simulation and Animation.**

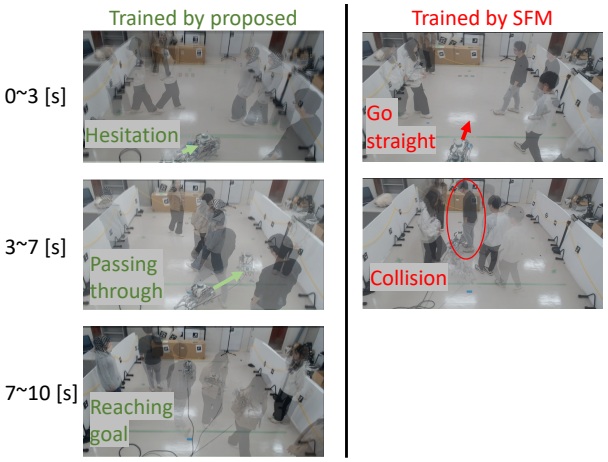


Fig. 1: Comparing robot navigation policies trained by different crowd models in the real world. (The experiment was conducted under ethical approval from the research ethics committee of the Graduate School of Frontier Sciences, University of Tokyo.)

I. Introduction

ROBOT navigation in human-populated environments has gained wide interest in recent years. In particular, attempts at applications such as unmanned patrolling, delivery, and cleaning using mobile robots are growing fast. To achieve these applications, safe and efficient robot navigation among a human crowd, known as "crowd navigation", is essential [1]. Although mobile robot navigation research has a long history, crowd navigation remains a challenge today since humans are also intelligent decision makers, exhibiting diverse behavior patterns [2]. A robust navigation policy that achieves harmonious human-robot cooperation with diverse human behaviors is desired.

To address such a problem, early rule-based methods such as the Social Force Model (SFM) require hyperparameter tuning for different crowd behaviors. In recent years, Deep Reinforcement Learning (DRL)-based methods that train navigation policies by self-exploration in simulation environments, have achieved great progress in crowd navigation [1], [3], [4]. However, crowd simulations used by these works lack diversity, which results in two limitations: (1) trained policies overfit some specific human behavioral patterns and have poor generalizability to unseen scenarios in practice; (2) trained policies get a poor sense of human uncertainty, which leads to risky movements.

Our goal is to develop a diversity-aware crowd model that reproduces diverse crowd navigation behaviors and provides enough sample diversity to train a robust robot navigation policy. In particular, we aim to capture the unpredictable nature of human behavior that leads to the uncertainty in human paths and is often described as the multi-modality of human movement [5], [6]. Though there are many complicated crowd navigation behaviors that exist for further study [7], such as lane-formation, grouping, and sudden goal-changing, this work considers reproducing diverse human local collision avoidance behavior as the first step since it largely affects robot's local plan. For this objective, many methods based on Generative Adversarial Imitation Learning (GAIL) have been proposed [5], [6], which learn from pre-collected data and generate diverse (multi-modal) crowd behaviors. However, generalization of these methods rely on highly representative datasets that are usually not available for unseen scenarios [8]. On the contrary, RL-based methods can learn to generate crowd models representing unseen scenarios without pre-collected datasets. Panayiotou *et al.* [9] employed a reward-

Manuscript received: January 6, 2025; Revised March 23, 2025; Accepted April 16, 2025.

This paper was recommended for publication by Editor Sven Behnke upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by JSPS KAKENHI under Grant 23KJ0580.

1: Jiaxu Wu, Yusheng Wang, Tong Chen, Yongdong Wang, Qi An and Atsushi Yamashita are with The University of Tokyo. wujiaxu@robot.t.u-tokyo.ac.jp

2: Jun Jiang is with Woven by Toyota jiangjun0105@gmail.com (work is done out of working hours).

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

weight-conditioned policy (CCP) to capture different human behaviors by concurrently training with different mixtures of atomic rewards. Although it reproduces many macroscopic crowd behaviors, such as goal-seeking and grouping, we find that it obtains less diversity in fine-grained behavior patterns in goal-seeking with collision avoidance (which is required by simulation for training the robot), such as different clearances to others or different side preferences. This is due to insufficient for latent space exploration since CCP uses policy entropy only [10], [11].

In this work, we propose a novel RL-based diversity-aware crowd model that employs Variational Exploration (VE) with mutual information (MI)-based auxiliary reward, allowing for more fine-grained behavioral pattern discovery during RL training. A latent-conditioned policy using a control code $z \sim P(z)$ is learned, generating distinct behavior per z, given the same situations. Although MI-based auxiliary reward is used in many related fields, such as skill discovery [11] and zero-shot human-AI coordination [12], the proposed method has two differences from them by considering the specification of the crowd simulation task. First, to generate near-optimal behaviors that mimic ordinary people, we employ the Constrained Partially Observable Markov Decision Process (Constrained POMDP) to constrain the optimality of the crowd model. Second, while previous works considered single-agent settings, crowd simulation has a multi-agent nature. We found that when multiple agents perform VE concurrently, training suffers from instability, resulting in policies with poor goal-seeking performance. To counter this problem, we perform VE in a Centralized Training Decentralized Execution (CTDE) paradigm.

Major contributions of this article are listed below.

- We propose a novel crowd model that realizes diverse and near-optimal crowd behaviors by a novel RL framework that integrated Constrainted VE (Section III-C) and CTDE paradigm (Section III-D).
- We collect a novel real-world dataset and perform a crowd simulation evaluation (Section IV-A1).
- We address robust robot navigation using the proposed diversity-aware crowd model and perform experiments for robot navigation in both simulation (Section IV-A2) and the real world (Section V).

II. Related Works

As a model-based method, [2] employed personality trait to represent diversity in human behaviors, in which different sets of ORCA parameters were designed to generate trajectories corresponding to different personalities. Similarly, an extended SFM is proposed in [13] where physique and mentality coefficients are proposed to modify the human self-driving force. Since this modification violates the symmetry assumption made by the SFM, careful engineering is required to obtain collision-free paths, therefore limiting the diversity of the simulation.

Data-driven methods learn human navigation policies or trajectory models from human trajectory datasets. Inverse Reinforcement Learning (IRL)-based methods learn human navigation policies by searching human reward functions using human trajectory feature matching [14]. GAIL-based methods have been popular in recent years. Zhou *et al.* employed an MI-based objective function to mimic diverse human decision-making behaviors exhibited in a dataset. Charalambous *et al.* trained a novelty detector as a discriminator to reject unrealistic behaviors generated by human policy [15]. Chen *et al.* employed the Diffusion model to capture multi-modal feature distribution in trajectory datasets [6]. The trained generator could generate various human trajectory patterns while achieving considerable diversity in crowd simulation. However, data-driven methods require highly representative datasets to make generalizations toward unseen scenarios, when such datasets are difficult to obtain. Another work closely related to this article is proposed by Ling *et al.*, in which a crowd model based on GAIL was learned from a dataset and applied to train the robot navigation policy [16]. However, this work did not consider diversity in human behaviors.

RL-based methods have been proposed in recent years. A parametric policy, conditioned by random human speed limits and desired speeds, is proposed by [17]. Though these methods can simulate heterogeneous crowds by setting different desired speeds to humans, it has not tackled multi-modality in human behaviors. The randomized reward function proposed by [9] is a close work for multi-modality in human behaviors, in which weights for different reward terms are randomized during training, and a conditional policy taking those reward weights as a condition vector is proposed. However, different reward terms compete with each other during training, resulting in a poor exploration of different behaviors. MAVEN uses VE under the multi-agent setting to encourage exploration [18]. However, it tackles only pure cooperation problems, but crowd simulation is mixed-motive (involving both cooperation and competition between humans).

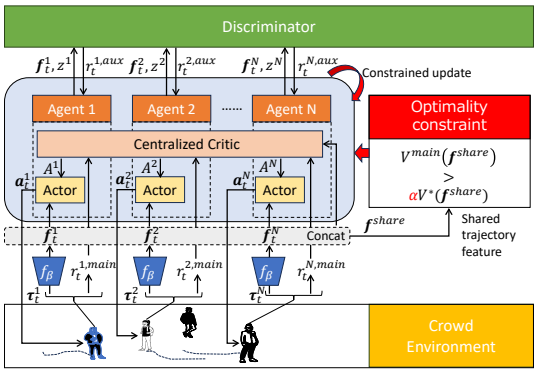
Although most of the previous works, as well as in this work, Human-Robot Interaction (HRI) is simplified, humans are assumed to have full attention on other agents [7], and they treat robots as the same as other humans [3], [16], this may not be true in the real world. HRI may also depend on how much attention humans pay to the robot [7], robot control policies [19], as well as human trust in the robot [20]. We plan to quantify and incorporate those factors into our crowd model in future work to further improve the reality and diversity of the simulated crowd.

III. Method

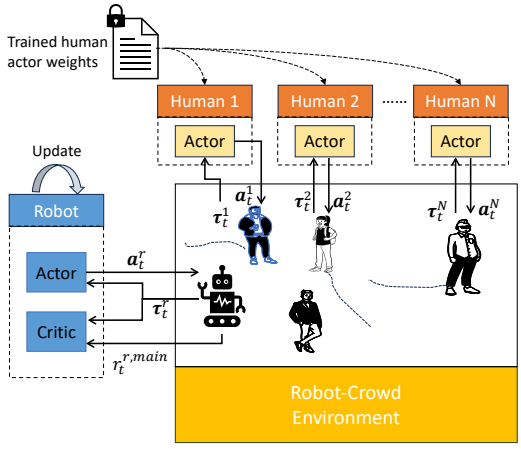
A. Overview

We realize diversity-aware crowd simulation using a latent-conditioned policy trained by a novel RL framework integrated with an MI-based auxiliary reward, optimality constraint, and centralized critic (Fig. 2). Previous work on diverse agent behavior using the RL framework with MI-based auxiliary rewards [11], [12] is difficult to train for scenarios with multiple humans due to their following a single-agent RL diagram and is also difficult to constrain the optimality of simulated humans close to the real world since they use fixed weights to combine task and auxiliary rewards. We address

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Centralized Training



(b) Decentralized Execution

Fig. 2: Overview of the proposed method. Human navigation policy is trained with MI-based auxiliary reward r^{aux} provided by the discriminator at first. Then, crowd simulation based on the trained human policy will be executed for training robot navigation policies.

the difficulty of training by introducing a centralized critic to the RL framework, forming a Centralized Training and Decentralized Execution diagram for our crowd simulation. The RL framework is further integrated with an optimality constraint to obtain near-optimal human navigation close to the real-world human.

B. Preliminary

Navigation tasks for both humans and robots can be formulated as a POMDP. Humans and robots are each referred to as a type of "agent" in the rest of this article.

1) Observation Space: The observation $\mathbf{o} = [\mathbf{o}^h, \Phi]$ contains the internal state of the agent $\mathbf{o}^h$, and a series of 720 depth sensing rays evenly spaced within a field of view (human: 110°, robot: 360°) $\Phi \in \mathbb{R}_0^{+,720}$, where the internal state includes agent's distance and relative heading to its goal ($d^g \in \mathbb{R}_0^+$ and $h^g \in [-\pi, \pi]$), velocity $\mathbf{v} \in \mathbb{R}^2$, preferred speed $v_{pref} \in \mathbb{R}_0^+$, and size $r \in \mathbb{R}_0^+$.

2) State Space: We consider the agent's state $\mathbf{s}_t = f_\beta(\tau_t)$ at time step t as the trajectory feature of the agent, which is extracted from the agent's trajectory $\tau_t = \mathbf{o}_{0:t}$ using a function f_β parameterized by a set of learnable parameters β. The agent state summarizes the temporal information, such as the motion of the surrounding humans and its own progress to the goal, from the historical observations of the agent.

3) Action Space: We adopt the unicycle model to drive humans and robots so that $\mathbf{a}_t = [v, w]$, where v denotes the speed command and w denotes the rotation command. The robot is controlled by a learnable control policy π_ϕ^r that takes partial observation $\mathbf{o}_t$ as input and outputs the velocity command $\mathbf{a}_t^r = \pi_\psi^r(\tau_t)$ and human is controlled by a latent-conditioned policy $\mathbf{a}_t^h = \pi_\phi^h(\tau_t, z)$ with the control code z.

4) Reward Function: We adopt a popular reward design from previous work [3],

$$r^{main}(\mathbf{o}_t, \mathbf{a}_t) = \begin{cases} r_g & \text{(reached goal)} \\ r_c & \text{(collided)} \\ \alpha_1 r_p + \alpha_2 r_d & \text{(else)} \end{cases}, \quad (1)$$

where α_1, α_2 are weights and r_p is the potential-based reward shaping function: $r_p(\mathbf{o}_{t-1}, \mathbf{o}_t) = d^g(\mathbf{o}_{t-1}) - d^g(\mathbf{o}_t)$. Since we introduce an auxiliary reward in the next section, we refer this reward function as the main reward r^{main}. $r_d(\mathbf{o}_t)$ is the uncomfortable penalty for the robot only, for getting too close to humans

$$r_d(\mathbf{o}_t) = \begin{cases} -d^{comf} + d^{min} & (d^{min} < d^{comf}) \\ 0 & \text{(else)} \end{cases}, \quad (2)$$

where d^{comf} refers to uncomfortable distance threshold. d^{min} is the minimum distance between the robot and all humans. $[r_g, r_c, d^{comf}, \alpha_1, \alpha_2]$ are set to $[10, -20, 1.0, 2, 0.2]$ for the robot, basically following the previous work [3] except the uncomfortable distance is set larger for larger separation distance between the robot and the human. For humans, the reward parameters are set to $[10, -20, 0, 2, 0]$ since the main reward for the human only considers goal-reaching and collision-free for obtaining more behavioral diversity.

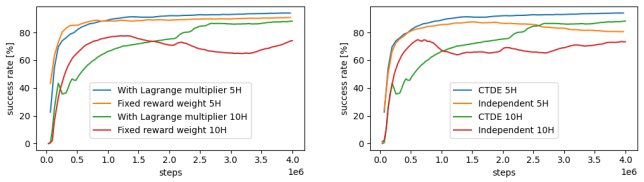
5) Problem Formulation: Based on the above preliminaries, our final goal is to realize a robust robot navigation policy $\mathbf{a}_t^r = \pi_\phi^r(\tau_t)$ by RL training with a diversity-aware crowd model for unseen scenarios. Our objective is to realize such diversity-aware crowd model by deriving a latent-conditioned human navigation policy $\mathbf{a}_t^h = \pi_\phi^h(\tau_t, z)$ without pre-collected data, which generates action commands that follow diverse behavior patterns corresponding to the control code z while maximizing cumulative reward r^{main} to a certain extent that approximates the optimality of real-world humans.

C. The Learning of Diverse Near-optimal Human Navigation Behaviors using Constrained Variational Exploration

To realize multiple near-optimal crowd behaviors given the same human initial states and goals, we propose a constrained VE following the previous Constrained Markov Decision Making formulation [21], in which a MI-based auxiliary reward is maximized given an optimality constraint on the expected cumulative main reward $R^{main} = \mathbb{E}[\Sigma_{t=0}^T \gamma^t r_t^{main}]$.

$$\pi^{h,*} = \underset{\pi^h}{\text{argmax}} \Sigma_{t=1}^T I(\mathbf{s}_t^h, z) \; s.t. V^{main,\pi^h}(\mathbf{s}) \geq \alpha V^{main,\pi*}(\mathbf{s}), \quad (3)$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Effect of Constrained POMDP. (b) Effect of CTDE

Fig. 3: Effectiveness of key components in the proposed RL framework. The validation curves, representing the navigation success rate during training, compare the performance of the proposed method with and without Constrained POMDP and CTDE in 5-human (5H) and 10-human (10H) circle crossing scenarios.

The MI-based auxiliary reward $r_t^{aux} = I(\mathbf{s}_t^h, z)$ is defined as the MI between a known distribution of control code z and an individual human state $\mathbf{s}_t^h$ appearing in the trajectory τ, where $\Sigma_{t=1}^{T} I(\mathbf{s}_t^h, z)$ is a lower bound for the MI between the distribution of z and the full human trajectory in the episode $\tau_{0:T}$ generated by the human policy π^h conditioned by z [11]. V^{main,π^h} represents the value function that approximates the cumulative main reward following π^h (referred to as the main value), and $V^{main,\pi*}$ represents the main value by following an optimal policy, which maximizes the cumulative main reward. The constraint in Eq. 3 requires the human policy to seek diverse behavioral patterns while maintaining goal-seeking and collision-free. $\alpha \in [0,1]$ denotes the relaxation coefficient.

The control code z is drawn uniformly from a discrete set $\{0,1,...N-1\}$ where N denotes the size of the set and is also referred to as the class number. The auxiliary reward encourages exploration, producing distinct trajectory patterns for different z values. Therefore, a larger range of z enables more diverse human navigation behavior. It is also worth noticing that static objects and other humans' movements constrain the trajectory pattern one can take, so that for a certain scenario, there should be an upper bound for the effective range of z. The calculation of MI-based auxiliary reward is aided by an extra discriminator $q_\varphi(z|\mathbf{s}_t^h)$ (Fig. 2a green box) which is parameterized by φ and is learned to approximate the posterior distribution $P(z|\mathbf{s}_t^h)$ by minimizing the Cross Entropy loss $L_{ce} = \frac{P(z|\mathbf{s}_t^h)}{q(z|\mathbf{s}_t^h)}$.

$$r^{aux} = \log N + \log q_\varphi(z|\mathbf{s}_t^h), \quad (4)$$

$$= \log N + \log q_\varphi(z|f_\beta(\tau_t)). \quad (5)$$

Since $f_\beta(\tau_t)$ represents the trajectory feature, the auxiliary reward will reward the agent when it follows trajectories with specific patterns corresponding to the control code z.

However, the policy tends to generate random behavior if there are no other reward signals or constraints. Different from previous work that uses fixed weight, to obtain near-optimal behaviors, we balance the main reward and the MI-based auxiliary reward with a dynamic weight, $r = \lambda r^{main} + r^{aux}$. λ is the Lagrange multiplier that is obtained by Proportional-Integral control using λ as control and $\alpha V^{main,\pi*}(\mathbf{s}) - V^{main,\pi^h}(\mathbf{s})$ as error feedback, for a smooth and proper update [22]. Intuitively, if the error is large, the PI controller will dramatically increase λ, and when the constraint is satisfied, the controller will reduce λ so that the VE can search for more diverse behavior. Fig. 3a shows that compared to the fixed weighting, the proposed method obtained more stable training and reached higher navigation success rates, especially when the scenario became crowded (10 humans).

D. Centralized Training Decentralized Execution

Unlike previous work [11] considering the single-agent setting, VE under multi-agent environments faces the instability problem caused by the difficulty in credit assignment and oscillation due to simultaneous exploration in multiple behavior latents. To counter this problem, we perform the proposed Constrained VE in CTDE paradigm and provide a practical implementation based on Heterogeneous Agent Proximal Policy Optimization (HAPPO). Different from the PPO for single-agent RL used in previous work [9], HAPPO uses centralized critics for the value approximation to address the credit assignment problem based on a novel advantage decomposition lemma [23].

The structure of our actor-critic model based on HAPPO is described as follows. To obtain the human state, the trajectory feature extraction f_β is implemented by a Recurrent Neural Network (RNN) with input embedding using a one-dimensional Convolutional Neural Network (1D-CNN)-based network structure from [24]. The human actions are then predicted by a 2-layered Multi-Layer Perceptron (MLP) using individual trajectory features, while value prediction is based on a shared feature $\mathbf{f}^{share}$ concatenated all humans' trajectory features (Fig. 2a gray box). We use MLP for the value prediction in this article, and more representative network structures like Graph Neural Network will be investigated in future work. The values and advantages of the main reward and the auxiliary reward are predicted separately. Instead of directly balancing the reward, this paper uses the Lagrange multiplier mentioned above to balance the advantages corresponding to different rewards ($A = \lambda A^{main} + A^{aux}$). Based on above, the actor-critic model is learned as follows.

$$\beta, \varphi, \phi \leftarrow L_{HAPPO}(\lambda A^{main} + A^{aux}, \mathbf{a}^h) + L_{ce}, \quad (6)$$

$$\xi^{main} \leftarrow \|(R^{main} - V^{main,\pi^h}(\mathbf{f}^{share}))\|_2, \quad (7)$$

$$\xi^{aux} \leftarrow \|(R^{aux} - V^{auxiliary,\pi^h}(\mathbf{f}^{share}))\|_2, \quad (8)$$

$$\lambda \leftarrow \mathrm{PI}(V^{main,\pi^*}(\mathbf{f}^{share}), V^{main,\pi^h}(\mathbf{f}^{share})), \quad (9)$$

where ξ^{main} and ξ^{aux} are weights for main and auxiliary value function. $\|\|_2$ denotes the L2 norm. PI denotes the Proportional-Integral control for adapting the Lagrange multiplier λ. Fig. 3b shows by using CTDE, the proposed method obtained more stability during training and reached higher navigation success rates (near-optimal behavior) compared to the results of using the single-agent RL diagram, especially when the scenario becomes crowded (10 humans).

IV. Simulation Experiments

To evaluate the effectiveness of the proposed crowd model:

- microscopic simulation accuracy compared to real-world crowds and diversity of the simulated trajectories were evaluated in Section IV-A1,
- significance on improving the robustness of robot navigation was tracked in Section IV-A2.

We selected baselines as follows. **Personality Trait ORCA (PT-ORCA)** [2]: tunes the parameters of ORCA to generate human trajectories corresponding to 3 different personalities. **Social Force Model (SFM)** [25]: non-diversity-aware ad-hoc method. Two models with different sets of hyperparameters: Aggressive model SFM-A $[v_0, \sigma] = [10, 0.3]$ and conservative model SFM-C $[v_0, \sigma] = [5, 1.5]$. **Configurable Crowd Profiles (CCP)** [9]: RL-based method that generates different human behaviors using randomization of the reward weights. **Parametric RL (P-RL)** [17]: non-diversity-aware RL method. Randomizes the preferred speed of the human as a parameter to select subspaces within the learned policy space. **SPDiff** [6]: data-driven crowd simulation using Diffusion model. Since we assume unseen scenarios without pre-collected data, we trained SPDiff using UCY dataset used in the original paper and tested it on our dataset.

A. Experimental Setup

1) Crowd Simulation Experiments: Since this work focuses on local navigation of mobile robots, crowd simulation evaluation experiments of the proposed method against the selected baselines were conducted in a 3.5×3.5 $[m]$ workspace. To evaluate the accuracy of the microscopic simulation, a comparison was performed between real-world human trajectories and simulated trajectories. A real-world dataset of the human navigation trajectory was collected in the 3.5×3.5 $[m]$ workspace using a motion capture system. 58 and 59 episodes were collected from each of the two-human interaction scenario and the three-human interaction scenario, respectively. For each episode, simulations were executed by each of the methods (the proposed and baselines) 20 times, given the human initial positions and destinations the same as in the real-world data. For all crowd models, simulations would be truncated at the moment when the real-world humans have reached their destinations. Since all RL-based methods (proposed, P-RL, and CCP) require the human maximum speed (also referred to as desired speed v_{pref}) to scale the outputs from the policy network and simulate human local navigation paths, we assessed the v_{pref} in each real-world episode and input it into the simulation.

The accuracy was assessed by ADE_{20} and FDE_{20}, which refer to the minimum Average Displacement Error (ADE) and Final Displacement Error (FDE) of 20 simulation results, widely used in previous works of multi-modal pedestrian trajectory prediction [26]. The number of collision (# collision) was also used since collision in real-world crowds is rare. The diversity of the simulations was qualitatively evaluated on the basis of the feature distributions of the generated trajectories. Two features were selected to characterize the generated trajectories: minimum separation distance from other humans (nn-dist) and winding angle (wd) that captures the side preference of the human, widely used in previous work [14], [27].

2) Robot Navigation Experiments: In this experiment, the contribution of the proposed crowd model to robust robot navigation in the crowd was evaluated. Two state-of-the-art actor-critic model: Distributed Multi-robot Collision Avoidance (DMCA) [24] and DRL-VO [4] were chosen as the training targets (both of them are open sourced). In navigation experiments, PT-ORCA was omitted from the baselines due to its low navigation success rate and high simulation error. SPDiff was also omitted since it requires pre-collected data and violates our assumption, as well as introducing a large number of collisions, which is considered non-realistic.

Two simulation workspaces were used: the circle cross in 8×8 $[m]$ room [1] and the square cross in 3.7×5.6 $[m]$ room [27]. Firstly, the proposed crowd model, P-RL, and CCP were trained in both workspaces with 5 humans. Robot navigation policies were then trained in the square-cross workspace with 4 humans driven by either the proposed crowd model or other baselines. To evaluate the robustness of the robot policies obtained, they were tested using 7 unseen scenarios formed from different workspaces and crowd models (square cross: proposed, P-RL, SFM-C, and SFM-A; circle cross: proposed, P-RL, SFM-A). In each scenario, robot navigation was tested for 500 episodes. The performance of robot navigation was evaluated in terms of the success rate (SR) and the average and standard deviation of the navigation time (NT) in total [1]. To further evaluate the robustness of the robot navigation across different scenarios, we also evaluate the standard deviation of SR (SR STD) across those 7 test scenarios. The results are shown in Section IV-B3.

3) Implementation Details: The proposed method was implemented based on the open-source Python repository provided by [23] [1]. In the crowd simulation evaluation experiments, we tested two different ranges of the control code ($z \in \{0,1,2\}$ and $z \in \{0,1,2,3,4,5\}$) to investigate the impact of the range of z on the simulation. The proposed method used $z \in \{0,1,2\}$ and $z \in \{0,1,2,3,4,5\}$ was denoted as Proposed(3C) and Proposed(6C) respectively. We selected these two ranges since previous domain studies [2] showed the existence of 6 different patterns in human movement during navigation and these patterns can be further summarized into 3 patterns according to PEN model [2]. We expected that z with a larger range would lead to larger diversity in human behavior. In the robot navigation simulation experiments, we focused on comparing the robustness of robot navigation trained by diversity-aware crowd simulation against non-diversity-aware crowd simulation so that only Proposed(3C) was used.

PT-ORCA was implemented based on the open-source Python library [2], and the hyperparameter was set to be the same as in the original article [2]. For a fair comparison, the observation space, action space, and human observation encoder of P-RL and CCP were modified to match the proposed method. In addition, for CCP, since this paper did not consider grouping and the human-object interaction, the corresponding reward weights were set to zero. The ranges of reward weights for reaching goals w_g and avoiding collisions

[1] https://github.com/PKU-MARL/HARL

[2] https://github.com/sybrenstuvel/Python-RVO2

w_{ca}, were set the same as the original article [9]. More details are available on the project home page.

All simulation experiments were conducted on a desktop PC with an AMD Ryzen Threadripper PRO 3955WX 16-Cores CPU and three Nvidia RTX 3090 GPUs. Training the proposed method to converge took $4,000,000$ steps and approximately 6 hours in wall-clock time.

B. Results

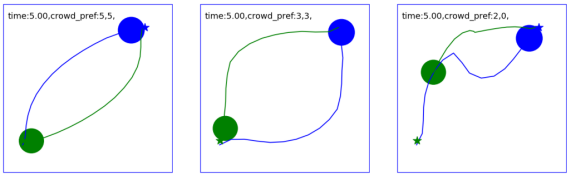
1) Crowd Simulation Accuracy: Table I shows the mean and standard deviation of ADE_{20} and FDE_{20} and the number of collisions of the proposed method and baselines. High FDE_{20} obtained by PT-ORCA implies that the ad hoc method failed to drive the simulated humans to their goals. In contrast, both Proposed(3C) and Proposed(6C) obtained smaller simulation error compared to all the baselines. On the other hand, the proposed method obtained more collision compared to the baselines except SPDiff. We consider the reason that mixing diverse behavior patterns in a scenario increases the difficulty of coordination between humans. SPDiff obtained a large number of collisions, and this may be due to the GAIL not being aware of collision.

TABLE I: Simulation evaluation results.

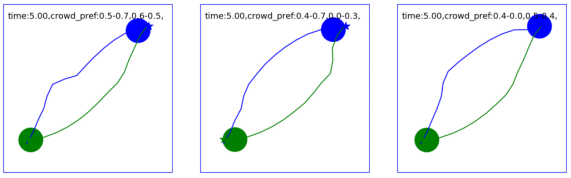
Metrics	ADE_{20}	FDE_{20}	# Collision
Proposed(6C)	0.257±0.104	0.082±0.034	20
Proposed(3C)	0.313±0.182	0.086±0.034	3
CCP	0.484±0.228	0.099±0.047	8
PT-ORCA	0.737±0.165	1.32±0.247	0
P-RL	0.617±0.246	0.191±0.123	0
SFM-C	0.614±0.211	0.378±0.388	0
SPDiff	0.415±0.155	0.076±0.125	150

2) Crowd Simulation Diversity: Simulated trajectories for a selected episode are shown in Fig. 4, in which the trajectories generated by CCP were almost the same given different sets of reward weights. In contrast, the proposed crowd model generated diverse interaction patterns between two humans given the same initial position and destination. Humans in the proposed model could have different minimum separation distances from others and different side preferences. Moreover, in addition to smooth coordination between two humans, the proposed method also generates suboptimal trajectories that reproduce the situation where two people fail to read the intentions of others.

Fig. 5a and Fig. 5b show the distribution of the minimum separation distance feature and the winding angle feature obtained by Proposd(3C), Proposed(6C), and CCP. For the minimum separation distance feature, both the real-world data and the simulation samples of the proposed method and the CCP concentrated around $0.5 \sim 1.0[m]$. Proposed(6C) obtained a wider distribution compared to Proposed(3C) which indicates more diverse human behavior patterns were generated using a larger range of z. The distribution of winding angle in the real-world dataset shows multi-modality, which indicates the existence of both right and left side preferences. The result implies that the proposed crowd model (both Proposed(3C) and (6C)) realized multi-modality in human behaviors, while CCP only reproduced right-side preference.

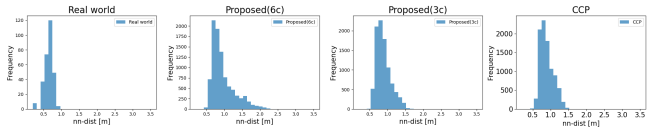


(a) Pass on the right (b) Large separation (c) Miss coordination

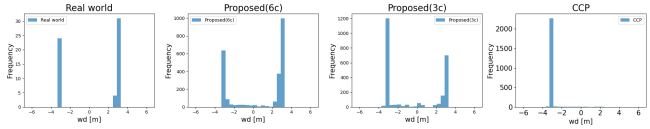


(d) (0.5,0.7),(0.6,0.5) (e) (0.4,0.7),(0.0,0.3) (f) (0.4,0.0),(0.5,0.4)

Fig. 4: Human trajectories simulated in $3.5 \times 3.5[m]$ workspace: Proposed(6C) (row 1) v.s. CCP (row 2). The green and blue circles denote two different humans whose destination is denoted by the star with the corresponding color. The sub-captions in row 2 represent different sets of reward weights.



(a) Minimum separation distance (nn-dist) in all scenarios



(b) Winding angle (wd) in the two-human interaction scenario

Fig. 5: Simulation evaluation results in 3.5×3.5 $[m]$ workspace with $2 \sim 3$ humans: trajectory feature distributions. The left-most column shows the distribution in the real world. Column 2: Proposed(6C), 3: Proposed(3C), 4: CCP.

3) Simulation Results of Robot Navigation: Table II shows results of robot navigation evaluation. For the overall success rate (column 2), the proposed method outperformed all baselines for both navigation methods (DMCA and DRL-VO), and the proportion Z-test indicates 99.9% confidence intervals between the proposed method and all baselines. Though for DMCA the SR only improved by 1.7%, this improvement can also be explained as the proposed method reduced 30% failure given 3500 trials. Since mobile robot navigation among humans is a failure-sensitive task (collision with humans probably causes injuries), we consider this reduction in failure to be meaningful for mobile robot safety. In addition, for both navigation methods, the proposed method obtained the smallest standard deviation of SR across the 7 unseen test scenarios (column 3). The results also show that DMCA obtained significantly higher SR compared with DRL-VO. We consider the reason is that the hyperparameter of DRL-VO was

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE II: Navigation evaluation results (DMCA/DRL-VO).

Metrics	SR	SR STD	NT [s]
Proposed(3C)	0.961/0.467	0.033/0.143	$6.67 \pm 2.73/6.63 \pm 2.48$
CCP	0.929/0.418	0.056/0.289	$6.63 \pm 2.29/6.90 \pm 1.78$
P-RL	0.944/0.389	0.057/0.177	$6.60 \pm 2.52/5.97 \pm 2.13$
SFM-C	0.910/0.215	0.042/0.210	$6.57 \pm 2.46/7.46 \pm 2.76$

adjusted for large scenarios with more free space (e.g. 20×10), so it struggles when the test scenario becomes smaller and is constrained by walls. Moreover, for DMCA, the proposed method achieved navigation times comparable to baseline methods with increments less than $0.1[s]$ (column 4). Although the proposed method obtained longer NT compared with P-RL ($+0.66s$) in the results of DRL-VO, the proposed method improved SR by 20%. These indicate a favorable trade-off between efficiency and safety, where the slight increase in time is justified by a substantial enhancement in success rate.

The above results imply that the proposed method can obtain more robust navigation policies in unseen scenarios in terms of success rate while allowing for a desirable trade-off in safety and efficiency. At the same time, properly selecting navigation method for the target scenario will further boost the performance.

V. Real World Robot Navigation Experiments

A. Experiment Setup

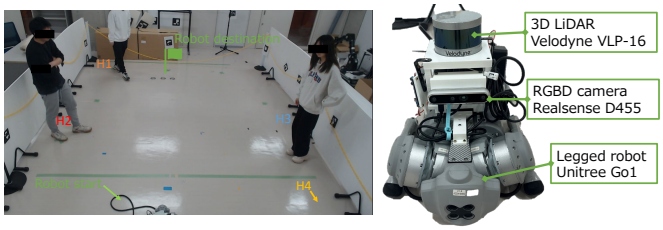


Fig. 6: Setup of the real-world experiments. The left picture illustrates the 3.7×5.6 $[m]$ square-cross scenario in the real-world setting. Humans started from all directions of the robot, right, left, front, and behind, and were denoted by H1, H2, H3, and H4.

The 3.7×5.6 $[m]$ square cross scenario was instantiated in a lab workspace, deploying a legged robot (Unitree go1 Edu) equipped with Velodyne VLP-16 LiDAR and Realsense D455 RGB-D camera to navigate next to 4 human participants (Fig. 6). Robot observations were obtained by self-localization based on RGBD odometry [28] and down-sampling of the LiDAR scan.

Two versions of the proposed crowd model, Proposed(3C): $z \in \{0,1,2\}$ and Proposed(24C): $z \in \{0,1,...23\}$, and the SFM-C crowd model were used to train three different robot navigation policies in the simulation with the same scenario, respectively. The trained policies were deployed on a notebook PC with an Intel Corei9-9980HK CPU and a Nvidia RTX 2080 Mobile GPU to control the robot.

36 trials of the scenario were executed in which the robot navigated with three different policies (12 trials per policy). In each trial, the robot started from the bottom-left area and navigated toward either the top-right corner of the workspace or the middle of the top boundary line (shown as the green flag in Fig. 6). 4 participants were told to navigate toward the destinations designated by the square-cross scenario. To ensure the diverse human behaviors in the experiment, we totally hired 4 women and 2 men, and we randomly picked 4 of them in each experimental trial and asked them to start from different starting positions in different trials. We also asked them to take different reactions to others, including the robot, such as different side preferences and different aggressiveness. The robot and humans interacted in the middle of the workspace so that its ability to avoid dense crowds could be evaluated. The local human density under this setting was around 0.55 human$/m^2$ which covered 85% of the scene in real-world dataset (Zara [6]).

B. Results

Table III summarizes the results of the policies trained by the proposed crowd model and SFM. It was found that the SFM-trained policy tends to make overconfident movements towards humans. This may lead to collisions when the nearby human also makes aggressive movements as shown on the right side of Fig. 1. The policy trained by the Proposed(3C) also showed a similar trend, but the robot will slow down and steer when being close to humans.

In contrast, the policy trained by Proposed(24C) usually makes hesitations before passing through dense crowds. The left side of Fig. 1 shows one example obtained by Proposed(24C), in which the robot was staying in place and attempting different directions to find a clear path. After a clear and safe path was found, the robot started to move and successfully reached the goal. However, the policy preferred to navigate near the wall and therefore erroneous self-localization and inaccurate or delayed motion command execution led to more robot-wall collisions (Table III row 3).

TABLE III: Results of the real-world tests

Crowd models	Succeeded	Collided / Human	Collided / Wall
SFM-C	11	1	0
Proposed(3C)	12	0	0
Proposed(24C)	10	0	2

VI. Discussion

For crowd simulation, the low diversity in the results of CCP indicates that policy entropy in PPO is insufficient to explore the behavior latent space. In contrast, the proposed Constrained VE may give almost random trajectories high rewards as long as they are distinct from trajectories generated under other control codes, which largely encourages exploration. For robot navigation, the superior performance of the policy trained by the proposed model in unseen scenarios proved the benefit of simulation diversity in RL training. The robot movements in the real-world test demonstrate that our proposed method yields an uncertainty-aware navigation

policy. We consider the reason that the simulated humans driven by our model exhibit multi-modality in training, so the robot should always prepare for different human movements in the same situations.

VII. Conclusion and Future Work

In this article, for robust robot navigation in human-populated environments, we propose a novel diversity-aware crowd model that can simulate diverse near-optimal human navigation behaviors for unseen scenarios without pre-collected data. Diverse human behaviors are realized by Constrained VE in the CTDE paradigm. Experiments on crowd simulation and robot navigation demonstrate that the proposed method achieves diverse crowd behaviors and improves the robustness of robot navigation, outperforming previous work. In future work, real-world crowd data involving more humans and complicated scenarios will be collected to enable more thorough simulation evaluations of the proposed model. The impact of the range of the control code z on the final robot navigation performance will also be tackled. Moreover, to improve the realism in the simulation of highly crowded scenarios, human collective behavior, such as line formation, will be tackled in the future.

References

[1] C. Chen, Y. Liu, S. Kreiss and A. Alahi: "Crowd-Robot Interaction: Crowd-Aware Robot Navigation with Attention-Based Deep Reinforcement Learning", Proceedings of the 2019 IEEE International Conference on Robotics and Automation (ICRA), pp. 6015–6022, May 2019.

[2] S. J. Guy, S. Kim, M. C. Lin and D. Manocha: "Simulating Heterogeneous Crowd Behaviors Using Personality Trait Theory", Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 43–52, August 2011.

[3] S. Liu, P. Chang, W. Liang, N. Chakraborty and K. Driggs-Campbell: "Decentralized Structural-RNN for Robot Crowd Navigation with Deep Reinforcement Learning", Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 3517–3524, May 2021.

[4] Z. Xie and P. Dames: "DRL-VO: Learning to Navigate Through Crowded Dynamic Scenes Using Velocity Obstacles", IEEE Transactions on Robotics (TRO), Vol. 39, No. 4, pp. 2700-2719, August 2023.

[5] H. Zou, H. Su, S. Song and J. Zhu: "Understanding Human Behaviors in Crowds by Imitating the Decision-Making Process", Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18), pp. 7648–7655, April 2018.

[6] H. Chen, J. Ding, Y. Li, Y. Wang and X. P. Zhang: "Social Physics Informed Diffusion Model for Crowd Simulation", IEEE Transactions on Visualization and Computer Graphics (TVCG), Vol. 38, No. 1, pp. 474–482, March 2024.

[7] C. Mavrogiannis, F. Baldini, A. Wang, D. Zhao, P. Trautman, A. Steinfeld and J. Oh: "Core Challenges of Social Robot Navigation: A Survey", ACM Transactions on Human-Robot Interaction (THRI), Vol. 12, No. 3, April 2023.

[8] G. Qiao, H. Zhou, M. Kapadia, S. Yoon and V. Pavlovic: "Scenario Generalization of Data-Driven Imitation Models in Crowd Simulation", Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG), Article 36, pp. 1-11, October 2019.

[9] A. Panayiotou, T. Kyriakou, M. Lemonari, Y. Chrysanthou and P. Charalambous: "CCP: Configurable Crowd Profiles", ACM SIGGRAPH 2022 Conference Proceedings (SIGGRAPH), Article 53, pp. 1–10, August 2022.

[10] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck and P. Abbeel: "VIME: Variational Information Maximizing Exploration", Proceedings of the 30th Advances in Neural Information Processing Systems (NeurIPS), pp. 1117–1125, 2016.

[11] S. Kumar, A. Kumar, S. Levine, and C. Finn: "One Solution is Not All You Need: Few-Shot Extrapolation via Structured MaxEnt RL", Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), pp. 8198-8210, 2020.

[12] A. Szot, U. Jain, D. Batra, Z. Kira, R. Desai, and A. Rai,"Adaptive Coordination in Social Embodied Rearrangement", Proceedings of the 40th International Conference on Machine Learning (ICML), pp. 33365–33380, 2023

[13] W. Wu, M. Chen, J. Li, B. Liu, and X. Zheng: "An Extended Social Force Model via Pedestrian Heterogeneity Affecting the Self-Driven Force", IEEE Transactions on Intelligent Transportation Systems (T-ITS), vol. 23, no. 7, pp. 7974–7986, 2022.

[14] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard: "Socially Compliant Mobile Robot Navigation via Inverse Reinforcement Learning", The International Journal of Robotics Research (IJRR), vol. 35, no. 11, pp. 1289–1307, 2016.

[15] P. Charalambous, J. Pettre, V. Vassiliades, Y. Chrysanthou, and N. Pelechano: "GREIL-Crowds: Crowd Simulation with Deep Reinforcement Learning and Examples", ACM Transactions on Graphics (TOG), vol. 42, no. 4, July 2023.

[16] B. Ling, Y. Lyu, D. Li, G. Gao, Y. Shi, X. Xu and W. Wu: "SocialGAIL: Faithful Crowd Simulation for Social Robot Navigation", Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 16873–16880, May 2024.

[17] K. Hu, B. Haworth, G. Berseth, V. Pavlovic, P. Faloutsos and M. Kapadia: "Heterogeneous Crowd Simulation Using Parametric Reinforcement Learning", IEEE Transactions on Visualization and Computer Graphics (TVCG), Vol. 29, No. 4, pp. 2036–2052, April 2023.

[18] A. Mahajan, T. Rashid, M. Samvelyan and S. Whiteson: "MAVEN: Multi-Agent Variational Exploration", Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS), pp. 7611–7622, 2019.

[19] A. Day and I. Karamouzas: "A Study in ZUCKER: Insights on Interactions Between Humans and Small Service Robots", IEEE Robotics and Automation Letters (RAL), Vol. 9, No. 3, pp. 2471–2478, March 2024.

[20] M. Chen, S. Nikolaidis, H.-C. Soh, D. Hsu and S. Srinivasa: "Planning with Trust for Human-Robot Collaboration", Proceedings of the 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 307–315, March 2018.

[21] T. Zahavy, Y. Schroecker, F. F. Behbahani, K. Baumli, S. Flennerhag, S. Hou and S. Singh: "Discovering Policies with DOMiNo: Diversity Optimization Maintaining Near Optimality", Proceedings of the International Conference on Learning Representations (ICLR), 2023.

[22] A. Stooke, J. Achiam and P. Abbeel: "Responsive Safety in Reinforcement Learning by PID Lagrangian Methods", Proceedings of the International Conference on Machine Learning (ICML), pp. 9133–9143, 2020.

[23] Y. Zhong, J. G. Kuba, X. Feng, S. Hu, J. Ji and Y. Yang: "Heterogeneous-Agent Reinforcement Learning", Journal of Machine Learning Research (JMLR), Vol. 25, No. 32, pp. 1–67, 2024.

[24] T. Fan, P. Long, W. Liu and J. Pan: "Distributed Multi-Robot Collision Avoidance via Deep Reinforcement Learning for Navigation in Complex Scenarios", The International Journal of Robotics Research (IJRR), Vol. 39, No. 7, pp. 856–892, 2020.

[25] D. Helbing and P. Molnár: "Social Force Model for Pedestrian Dynamics", Physical Review E, Vol. 51, No. 5, pp. 4282–4286, May 1995.

[26] T. Salzmann, B. Ivanovic, P. Chakravarty and M. Pavone: "Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data", Proceedings of the European Conference on Computer Vision (ECCV), pp. 683–700, August 2020.

[27] C. Mavrogiannis, K. Balasubramanian, S. Poddar, A. Gandra and S. S. Srinivasa: "Winding Through: Crowd Navigation via Topological Invariance", IEEE Robotics and Automation Letters (RAL), Vol. 8, No. 1, pp. 121–128, 2023.

[28] M. Labb´e and F. Michaud, "RTAB-Map as an Open-Source Lidar and Visual SLAM Library for Large-Scale and Long-Term Online Operation," Journal of Field Robotics (JFR), vol. 36, no. 2, pp. 416–446, 2019.