Towards Generalizable Safety in Crowd Navigation via Conformal Uncertainty Handling

Jianpeng Yao, Xiaopan Zhang*, Yu Xia*, Zejin Wang, Amit K. Roy-Chowdhury, and Jiachen Li†

University of California, Riverside
{jyao073, jiachen.li}@ucr.edu

Abstract: Mobile robots navigating in crowds trained using reinforcement learning are known to suffer performance degradation when faced with out-of-distribution scenarios. We propose that by properly accounting for the uncertainties of pedestrians, a robot can learn safe navigation policies that are robust to distribution shifts. Our method augments agent observations with prediction uncertainty estimates generated by adaptive conformal inference, and it uses these estimates to guide the agent's behavior through constrained reinforcement learning. The system helps regulate the agent's actions and enables it to adapt to distribution shifts. In the in-distribution setting, our approach achieves a 96.93% success rate, which is over 8.80% higher than the previous state-of-the-art baselines with over 3.72 times fewer collisions and 2.43 times fewer intrusions into ground-truth human future trajectories. In three out-of-distribution scenarios, our method shows much stronger robustness when facing distribution shifts in velocity variations, policy changes, and transitions from individual to group dynamics. We deploy our method on a real robot, and experiments show that the robot makes safe and robust decisions when interacting with both sparse and dense crowds. Our code and videos are available on https://gen-safe-nav.github.io/.

Keywords: Crowd Navigation, Reinforcement Learning, Distribution Shift

1 Introduction

Safely navigating among crowds is fundamental to a future where robots work closely with humans [1, 2, 3, 4]. Various solutions have been studied, including rule-based methods [5, 6], optimization-based planners [7, 8, 9, 10], reinforcement learning (RL) approaches [11, 12, 13, 14], and hybrid systems that track a reference path from an optimizer while a learned policy makes necessary adaptations [15, 16]. For RL approaches, a single millisecond-level forward pass can generate decisions, making them faster and more scalable than optimization- or rule-based planners. Yet while they excel in in-distribution settings, their performance drops sharply in out-of-distribution (OOD) environments [17, 18]. This indicates a tendency towards overfitting and difficulty in generalizing to diverse crowd dynamics [19].

Moreover, previous studies have shown that incorporating predictions into observations, thus forming prediction-augmented observations [20, 21, 22, 23, 24, 25, 26], can explicitly represent human intentions [27, 28] and aid in robot decision making. However, this practice may exacerbate overfitting issues, as human dynamics in real-world scenarios are inherently complex and difficult to fully capture in simulation environments or datasets, and trajectory predictions fitted for one certain dynamics face generalization challenges to other dynamics. Inaccurate predictions can severely mislead robot decisions, especially when robots rely heavily on them to determine actions. Due to

[*] Equal contribution; [†] Corresponding author

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

compounded overfitting from both predictions and the learned policy, robots specialized for these environments often fail to generalize to new scenarios with different crowd dynamics. In other words, existing methods lack a systematic treatment of prediction errors to improve policy generalizability.

In this paper, we propose that by properly quantifying uncertainties in human trajectory predictions and incorporating the results into decision making algorithms, it is possible to alleviate overfitting in RL-based crowd navigation. Uncertainty serves as an indicator of the reliability of prediction hypotheses, reflecting both prediction errors and the sensitivity of prediction models to distribution shifts [29]. Therefore, encouraging the agent to account for uncertainty enables it to generate actions that are more robust to distribution shifts and resilient to incorrect assumptions about human dynamics. Specifically, we introduce a learning-based framework that explicitly reasons about prediction uncertainty. First, we apply adaptive conformal inference (ACI) [30, 31] to quantify the uncertainty of each predicted human trajectory with a prediction set that contains the true future position with a user-defined coverage probability. Unlike other conformal approaches [32], ACI updates its calibration online, so it can swiftly adapt when the underlying crowd dynamics shift. Second, we employ constrained reinforcement learning (CRL) to introduce effective controllability into the decision making system, using uncertainty estimates to guide both the learning process and the agent's behavior. Our system achieves the state-of-the-art (SOTA) performance in safety metrics and with much smaller performance drops in OOD settings. In the in-distribution setting, our system achieves an over 8.80% higher success rate than previous SOTA RL baselines, with more than 3.72 times fewer collisions and over 2.43 times fewer intrusions into pedestrian trajectories in in-distribution settings. In three different OOD test scenarios that introduce velocity shifts, policy shifts, and pedestrian grouping behavior, our method maintains a high success rate and low collision rate, while competing approaches degrade significantly. Finally, we deploy the learned policy on a real Mecanum-wheel robot. With only minor clipping and smoothing, the policy transfers directly from simulation and achieves safe navigation in both sparse and dense crowds.

2 Related Work

Crowd Navigation. Mobile robots are expected to interact with humans and complete various tasks, such as providing assistance [1]. Crowd navigation forms the foundation for performing most high-level tasks. Robots are required to navigate in crowds, where the challenge of modeling dynamic human behavior makes navigation challenging. It is crucial to capture the subtleties of human behavior, such as human intentions and interactions between agents [33, 34, 3, 35, 36, 37], and properly use them for effective robotic decisions. Deep RL offers a potentially viable solution to the challenging navigation task [10, 38]. Previous work on RL-based methods for social robots includes the capture of agent-agent interactions [39, 40] and the intentions of human agents [21, 41], incorporating these as predictions into RL policy networks. Our work focuses on alleviating performance degradation in OOD scenarios and proposes a system that works towards this challenge by effectively considering and handling uncertainties in human behavior predictions.

Planning Under Uncertainty. Trajectory planning under uncertainty has attracted increasing attention in the past decades. In optimization-based and search-based approaches, researchers have integrated uncertainty quantification from perception and prediction into various controllers [42, 43, 44, 45, 46]. These methods easily support adding constraints or safety shields, allowing explicit management of uncertainty. By contrast, guiding decision making in RL agents with uncertainty measures is less straightforward, since policy networks often behave as black boxes. Prior work has augmented agent observations with uncertainty estimates [47, 48] or post-processed policy outputs based on uncertainty to ensure safe behavior [49]. However, most approaches avoid integrating uncertainty guidance directly into the learning process and still suffer from OOD performance degradation similar to the methods without uncertainty measures. In our work, we propose that uncertainty-aware planning can effectively mitigate OOD performance degradation for RL-based crowd navigation and introduce a CRL framework that enhances controllability and leverages uncertainty estimates to effectively guide decision making, thereby improving robustness to prediction errors and OOD scenarios.

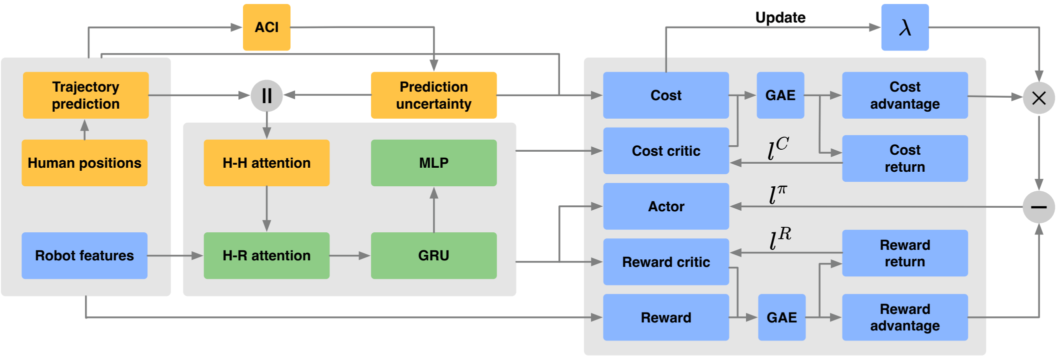

Figure 1: The overall pipeline of our method. We mark components related to humans in yellow, components related to physical information and decision making of the robot in blue, and fused features in green. We use ACI to quantify the prediction uncertainty of human trajectories and concatenate these metrics with predictions before inputting them into networks. The networks contain attention mechanisms for interactions between humans (H-H attention) and between humans and the ego robot (H-R attention). Prediction uncertainty combined with physical information is used for designing costs. For the CRL agent using PPO Lagrangian, the actor and reward critic share some layers while the cost critic uses a separate network. We adopt reward value loss l^R, action loss l^π, and cost value loss l^C for updating the agent.

3 Method

3.1 Problem Formulation

In our setting, we have H humans in the environment, each indexed by h, within an episode of horizon T. At each time step t, the positions of humans are represented as $\mathbf{p}_h(t)$. We predict K future steps (a larger K means more extended future predictions) for each human's future trajectories. The prediction point for the k-th prediction of the h-th human is denoted as $\mathbf{p}_{h,k}(t)$, where $1 \leq h \leq H$ and $1 \leq k \leq K$. We formulate the task as a constrained Markov decision process, where the CRL agent is provided with observations of the state S_t at each timestep, which consists of two main parts. The first part includes physical information: the current positions of humans and the robot, and other quantities about the robot's dynamics. The second part comprises post-processed features generated by models, such as the predicted human trajectories. We denote physical state components of ego information as $\mathbf{e}$, physical components of human information as $\mathbf{h}$, and components generated by models as $\mathbf{m}$. The complete state S_t is written as $S_t = [\mathbf{e}, \mathbf{h}, \mathbf{m}]$. The agent then generates action $A_t = (v_x, v_y)$ to control the moving speed of the robot according to S_t. After taking an action, the environment provides a reward R_t and a cost C_t. We aim to obtain an optimal policy $\pi(A_t \mid S_t)$ that maximizes rewards while satisfying cost constraints.

3.2 Method Overview

The overall pipeline of our method is illustrated in Fig. 1. First, we employ two different trajectory predictors: the constant velocity (CV) predictor [50] and the Gumbel social transformer (GST) predictor [51]. This is to demonstrate that our algorithm can adapt to both rule-based and learning-based predictors. Next, we use ACI to quantify prediction uncertainties and incorporate them into $\mathbf{m}$; subsequently, we adopt a policy network with a combined attention mechanism as proposed in [21] to process these uncertainties along with other features. Finally, we employ CRL to guide the agents' behavior using the uncertainty quantification results. Instead of directly applying constraints to the collision rate, we impose constraints on the cumulative intrusions of the robot into other agents' uncertainty areas. This approach provides behavior-level guidance and effectively addresses the issue of sparse constraint feedback, thereby improving upon previous methods that constrained result-oriented metrics such as the collision rate.

3.3 Rule-Based and Learning-Based Trajectory Prediction

Our system adapts to different prediction models, including both learning-based and rule-based trajectory predictors, and can mitigate the adverse effects of incorrect predictions on the subsequent decision making process. For the rule-based prediction model, we use the CV model to obtain simple and effective estimates of future human states by extrapolating the current states of human agents based on their velocities. For learning-based prediction models, we choose the GST predictor [51] as the learning-based prediction model, which is designed to address the challenges of partially detected pedestrians and redundant interaction modeling. GST is adapted for the efficient encoding of pedestrian features and demonstrates adaptability and robustness in high-density crowds.

3.4 Adaptive Conformal Inference for Quantifying Prediction Uncertainty

After obtaining the K-step future predictions, we quantify the prediction uncertainty using dynamically-tuned adaptive conformal inference (DtACI) [31], an ACI algorithm that adapts effectively to distribution shifts and thus serves as a good fit for online uncertainty estimation in social navigation. For each prediction step of each pedestrian, we run M prediction error estimators simultaneously. At time step t, we calculate the actual prediction error $\delta_{h,k}$ between the current position and the predicted position made at time step $t-k$ for the k-th prediction step of the h-th human as $\delta_{h,k}(t) = \|\mathbf{p}_h(t) - \mathbf{p}_{h,k}(t-k)\|_2$, where $\mathbf{p}_h(t)$ is the position of the h-th human at time t, $\mathbf{p}_{h,k}(t-k)$ is the k-step predicted position of the h-th human made at time $t-k$, and $\delta_{h,k}(t)$ is calculated as the L2 norm of the difference of the two values. We then update the estimated error generated by the m-th estimator for the h-th human as

$$\hat{\delta}_{h,k}^{(m)}(t) = \hat{\delta}_{h,k}^{(m)}(t-1) \; - \; \gamma^{(m)}\big(\alpha - \mathrm{err}_{h,k}^{(m)}(t)\big), \tag{1}$$

where $\hat{\delta}_{h,k}^{(m)}$ represents the estimated prediction error of the m-th estimator corresponding to a k-step ahead prediction for the h-th human, $\gamma^{(m)}$ is the learning rate of the m-th prediction error estimator for all humans and predictions, α is the coverage parameter, and

$$\mathrm{err}_{h,k}^{(m)}(t) := \begin{cases} 1, & \text{if } \hat{\delta}_{h,k}^{(m)}(t-1) < \delta_{h,k}(t), \\ 0, & \text{if } \hat{\delta}_{h,k}^{(m)}(t-1) \geq \delta_{h,k}(t). \end{cases} \tag{2}$$

According to quantile regression [31], $\hat{\delta}_{h,k}^{(m)}$ converges to the $(1-\alpha)$ quantile of actual errors. Since we run M prediction error estimators with different learning rates simultaneously, for each error estimator, after taking in the actual prediction error $\delta_{h,k}$ and updating the estimated prediction error for the next step, we evaluate the errors of each estimator and update the probability distribution for choosing the next output prediction error estimator by

$$w_{h,k}^{(m)} \leftarrow (1-\sigma)\,\frac{w_{h,k}^{(m)}\exp\big(-\eta\,\ell(\delta_{h,k}, \delta_{h,k}^{(m)})\big)}{\sum_{j=1}^{M} w_{h,k}^{(j)}\exp\big(-\eta\,\ell(\delta_{h,k}, \delta_{h,k}^{(j)})\big)} \; + \; \frac{\sigma}{M}, \quad p_{h,k}^{(m)} \leftarrow \frac{w_{h,k}^{(m)}}{\sum_{j=1}^{M} w_{h,k}^{(j)}}, \tag{3}$$

where we ignore the explicit time-step notation t for simplicity, and use the arrow to indicate the replacement of values. The weight $w_{h,k}^{(m)}$ corresponds to the probability $p_{h,k}^{(m)}$ for the m-th prediction error estimator, σ and η are hyperparameters of DtACI for adjusting the speed at which weights change. $\ell(\delta_{h,k}, \delta_{h,k}^{(m)})$ is the pinball loss function used to measure the estimation error:

$$\ell(\delta_{h,k}, \delta_{h,k}^{(m)}) = \begin{cases} \alpha\left(\delta_{h,k} - \delta_{h,k}^{(m)}\right), & \text{if } \delta_{h,k} \geq \delta_{h,k}^{(m)}, \\ (\alpha - 1)\left(\delta_{h,k} - \delta_{h,k}^{(m)}\right), & \text{if } \delta_{h,k} < \delta_{h,k}^{(m)}. \end{cases} \tag{4}$$

Each time we estimate the prediction uncertainty, we treat $\hat{\delta}_{h,k}$ as a discrete random variable taking values in the set $\{\hat{\delta}_{h,k}^{(1)}, \ldots, \hat{\delta}_{h,k}^{(M)}\}$, with probability mass function taken values from $\{p_{h,k}^{(m)}\}$.

3.5 Policy Network Structure

Once we have obtained the trajectory prediction results and the corresponding prediction uncertainty, we concatenate the uncertainty quantification with the predicted trajectory before feeding it into the attention layers. This allows the RL agents to account for the prediction uncertainty in their decision making process, as shown in Fig. 1. The first block is the human-human attention (H-H attention in Fig. 1), which models each human as a separate node to capture interactions among humans. Next, we fuse the robot features (including velocity, heading, positions, and goal) into the attention blocks through human-robot attention (H-R attention) to obtain fused feature embeddings capturing the interactions between humans and the ego robot. We then process these fused embeddings with the robot features and concatenate them to form the GRU input, thereby capturing temporal information. Lastly, the final fused features are passed to the actor and critic networks for further processing. For more details about the policy network structure, please refer to [21].

3.6 Decision Making with Effective Uncertainty Handling

To enhance controllability and enable effective guidance under uncertainty, we adopt a physically meaningful cost and adjust its distribution based on a predefined cost limit using CRL, which is not achievable with traditional RL and reward shaping, as conventional rewards provide only an aggregate signal without any mechanism to ensure that safety-related components converge within a specified range. First, we design the safety critical area of pedestrians as a combination of a circular area around the human's position and an uncertainty area around K' ($K' \leq K$) steps of predictions. Since we have H human agents in the environment, the two parts of the area are defined as

$$D_i(\mathbf{p}_{\text{ego}}) = \{\mathbf{p}_{\text{ego}} : |\mathbf{p}_{\text{ego}} - \mathbf{p}| \leq r_i\}\,,\ \mathbf{p} \in P_i,\ i = 1, 2, \tag{5}$$

$$P_1 = \{\mathbf{p}_h\},\ P_2 = \{\mathbf{p}_{h,k}\},\ 1 \leq h \leq H,\ 1 \leq k \leq K', \tag{6}$$

$$r_1 = r_{\text{ego}} + r_h + r_{\text{comfort}}, \quad r_2 = r_{\text{ego}} + r_h + \hat{\delta}_{h,k}, \tag{7}$$

where D_1 is the subarea around the current positions of humans and D_2 is the subarea around the predicted positions of humans. If the current center position of the ego robot $\mathbf{p}_{\text{ego}}$ is in either D_1 or D_2, an intrusion occurs. For the computation, we consider the distance between the center positions of agents and prediction points. r_{ego} is the radius of the ego robot, r_h is the radius of the h-th human, r_{comfort} is the radius of the subarea around the current positions of humans, and $\hat{\delta}_{h,k}$ is the prediction uncertainty generated by DtACI for the k-th prediction point of the h-th human.

At each time step t, we iterate through all cost areas of all humans and calculate the maximum intrusion, denoted as $d_{\text{intru},t}$. For an episode with a horizon of T, we have

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{T} R_t(S_t, A_t)\right] \quad \text{s.t.} \quad \mathbb{E}_{\pi}\left[\sum_{t=0}^{T} d_{\text{intru},t}\right] = \tilde{d}, \tag{8}$$

where $\tilde{d}$ is a pre-defined threshold.

We formulate the cost C_t using the intrusions as $C_t(S_t, A_t) = \mu d_{\text{intru},t}$, where μ is a constant. Our reward includes three components: R_{success}, $R_{\text{collision}}$, and $R_{\text{potential}}$. R_{success} is the reward for robot reaches the goal, $R_{\text{collision}}$ is for the robot collides with pedestrians, and $R_{\text{potential}}$ provides a dense reward that drives the ego robot to approach the goal, proportional to the distance the ego robot approaches the goal compared to the previous time step [21].

In our work, we use the PPO Lagrangian [52] for optimization. We set up two critics to compute the state value for reward and the state value for cost. The loss functions for the critics are defined as

$$l_t^R = c_1(V_{\theta_1}^R(S_t) - V_t^{\text{targ},R})^2, \tag{9}$$

$$l_t^C = c_2(V_{\theta_2}^C(S_t) - V_t^{\text{targ},C})^2, \tag{10}$$

where c_1 and c_2 are constants, $V_{\theta_1}^R(S_t)$ and $V_{\theta_2}^C(S_t)$ are network-generated value estimates for reward and cost, respectively, and $V_t^{\text{targ},R}$ and $V_t^{\text{targ},C}$ are the corresponding target values.

As for the policy network, the action loss is similar to the form in PPO [53] except that we employ the combined advantage $\hat{A}'_t = \frac{\hat{A}_t^R - \lambda \hat{A}_t^C}{1+\lambda}$. The action loss function can then be written as

$$l_t^{\pi} = \hat{\mathbb{E}}_t\left[\min\left(r_t(\theta_3)\hat{A}'_t, \text{clip}\left(r_t(\theta_3), 1-\epsilon, 1+\epsilon\right)\hat{A}'_t\right)\right], \tag{11}$$

Table 1: In-Distribution Test Results

Methods	**SR↑**	**CR↓**	**TR↓**	**NT↓**	**PL↓**	**ITR↓**	**SD↑**
SF	15.60%	21.44%	62.96%	30.23	34.64	3.78%	0.42
ORCA	67.84%	27.52%	4.64%	22.80	19.74	**1.10**%	**0.50**
MPC	73.76%	25.52%	0.72%	19.09	20.88	11.91%	0.43
SafeCrowdNav	89.09±3.18%	10.91±3.18%	**0.00±0.00**%	13.06±0.34	**12.06±0.45**	14.09±0.91%	0.39±0.01
CrowdNav++	86.11±0.61%	13.81±0.57%	0.00±0.14%	14.96±1.07	20.68±0.88	6.61±0.98%	0.43±0.01
RL (w/o ACI)	92.67±1.51%	7.33±1.51%	**0.00±0.00**%	**12.89±0.10**	19.42±0.12	10.62±0.57%	0.39±0.01
RL (w/ ACI)	94.08±1.18%	5.92±1.18%	**0.00±0.00**%	13.35±0.26	19.88±0.25	8.81±0.67%	0.40±0.00
Ours (w/ CV)	96.03±1.14%	3.73±1.24%	0.24±0.24%	17.88±0.60	24.51±0.76	2.40±0.22%	0.45±0.00
Ours (w/ GST)	**96.93±0.68**%	**2.93±0.61**%	0.13±0.12%	17.54±0.86	24.27±0.85	2.72±0.16%	0.44±0.00

Table 2: Out-of-Distribution Test Results

Environments	Methods	**SR↑**	**CR↓**	**TR↓**	**NT↓**	**PL↓**	**ITR↓**	**SD↑**
Rushing Humans	SF	12.24%	19.12%	68.64%	32.06	36.15	5.31%	0.40
	ORCA	60.32%	34.96%	4.72%	23.41	19.84	**2.95** %	**0.48**
	MPC	49.76%	49.84%	0.40%	18.52	17.77	19.70%	0.39
	SafeCrowdNav	69.71±4.55%	30.29±4.55%	**0.00±0.00**%	13.47±0.40	**10.96±0.66**	20.20±1.10%	0.37±0.00
	CrowdNav++	73.17±1.24%	26.67±1.40%	0.16±0.16%	15.43±1.63	19.89±1.12	12.38±1.48	0.39±0.00
	RL (w/o ACI)	74.19±1.26%	25.81±1.26%	**0.00±0.00**%	**13.31±0.30**	18.20±0.10	18.23±0.31%	0.37±0.00
	RL (w/ ACI)	76.96±4.29%	23.04±4.29%	**0.00±0.00**%	14.13±0.32	19.04±0.14	15.84±1.14%	0.37±0.01
	Ours (w/ CV)	87.07±0.89%	**12.75±1.21**%	0.19±0.32%	18.74±0.31	24.57±0.66	5.29±0.20%	0.40±0.00
	Ours (w/ GST)	**87.17±4.14**%	**12.75±4.00**%	0.08±0.14%	18.32±1.01	24.04±0.85	6.82±1.36%	0.38±0.00
SF Pedestrian Model	SF	12.08%	6.72%	81.20%	29.76	40.56	1.60%	0.45
	ORCA	92.56%	4.88%	2.56%	22.36	21.91	**0.72**%	**0.48**
	MPC	89.76%	10.00%	0.24%	17.07	20.58	8.58%	0.41
	SafeCrowdNav	91.28±0.60%	8.72±0.60%	**0.00±0.00**%	**12.12±0.23**	**11.37±0.29**	10.94±0.40%	0.39±0.00
	CrowdNav++	92.48±1.36%	7.52±1.36%	**0.00±0.00**%	14.65±1.81	20.82±1.71	6.48±0.88%	0.41±0.00
	RL (w/o ACI)	95.68±0.89%	4.32±0.89%	**0.00±0.00**%	12.35±0.21	18.99±0.16	9.98±0.49%	0.39±0.00
	RL (w/ ACI)	97.41±0.81%	2.59±0.81%	**0.00±0.00**%	13.08±0.24	19.81±0.24	8.07±0.43%	0.40±0.01
	Ours (w/ CV)	98.48±0.92%	1.39±0.82%	0.13±0.12%	19.02±0.45	25.84±0.54	2.04±0.25%	0.43±0.01
	Ours (w/ GST)	**98.96±0.52**%	**1.04±0.52**%	**0.00±0.00**%	18.18±0.84	25.05±0.79	2.66±0.45%	0.42±0.00
Groups	SF	2.56%	97.44%	**0.00**%	**9.41**	**9.91**	23.59%	0.35
	ORCA	49.44%	46.72%	3.84	22.98	19.37	20.46%	0.39
	MPC	69.12%	29.76%	1.12%	20.48	22.78	24.08%	0.37
	SafeCrowdNav	75.73±2.01%	24.27±2.01%	**0.00±0.00**%	16.29±0.44	13.44±0.64	26.78±1.42%	0.35±0.01
	CrowdNav++	81.84±15.17%	18.13±15.20%	0.03±0.05%	17.28±1.30	24.95±3.60	9.05±6.54%	0.39±0.01
	RL (w/o ACI)	71.07±6.33%	28.93±6.33%	**0.00±0.00**%	14.87±0.67	21.35±1.08	20.17±2.07%	0.36±0.01
	RL (w/ ACI)	83.12±7.81%	16.88±7.81%	**0.00±0.00**%	15.70±0.39	23.53±1.30	12.67±4.83%	0.37±0.01
	Ours (w/ CV)	**94.51±4.20**%	**5.41±4.23**%	0.08±0.14%	21.16±1.32	30.61±1.98	**3.00±0.61**%	**0.43±0.01**
	Ours (w/ GST)	94.13±3.51%	5.71±3.31%	0.16±0.21%	20.19±0.69	29.69±0.81	3.55±1.21%	0.41±0.01

where $r_t(\theta_3)$ represents the change ratio between the updated and old policies, ϵ is the clip ratio, and $\hat{A}'_t$ represents the estimated value advantage function at time step t. Lastly, we update λ using gradient descent. The loss function [54] for updating λ is defined as $l^{\lambda}_t = -\lambda(\bar{C} - \tilde{d}_C)$, where $\bar{C}$ is the mean episodic cost and $\tilde{d}_C$ is the *cost limit* proportional to $\tilde{d}$.

4 Experiments

4.1 Simulation Settings, Evaluation Metrics, and Implementation Details

We adopt CrowdNav [39], one of the most widely adopted simulation environments for crowd navigation. The environment consists of 20 humans and one robot in a $12\,\mathrm{m} \times 12\,\mathrm{m}$ area, with randomized initial positions and goal locations. The robot has a radius of $0.2\,\mathrm{m}$, while human radii are randomly sampled between $0.3\,\mathrm{m}$ and $0.5\,\mathrm{m}$. The robot's maximum speed is set to $1.0\,\mathrm{m/s}$. We adopt the standard evaluation metrics [21], including success rate (SR), collision rate (CR), timeout rate (TR), navigation time (NT), path length (PL), intrusion time ratio (ITR), and social distance (SD). The detailed definitions can be found in Appendix D. For all of these metrics, SR reflects the effectiveness and CR evaluates the safety of algorithms. TR, NT, and PL measure the efficiency of the policy-generated paths, and ITR and SD measure the politeness of the paths. While different deployment contexts may prioritize different metrics, we emphasize SR and CR given their critical importance for safe robot deployment. We include our implementation details in Appendix E.

4.2 Baselines and Ablation Models

Our baselines include optimal reciprocal collision avoidance (ORCA) [5], social force (SF) [6], model-predictive control (MPC), SafeCrowdNav [23], and CrowdNav++ [21]. ORCA and SF are classic rule-based algorithms in obstacle avoidance, MPC is an optimization-based algorithm, and SafeCrowdNav and CrowdNav++ represent the previous SOTA RL algorithms in crowd navigation.

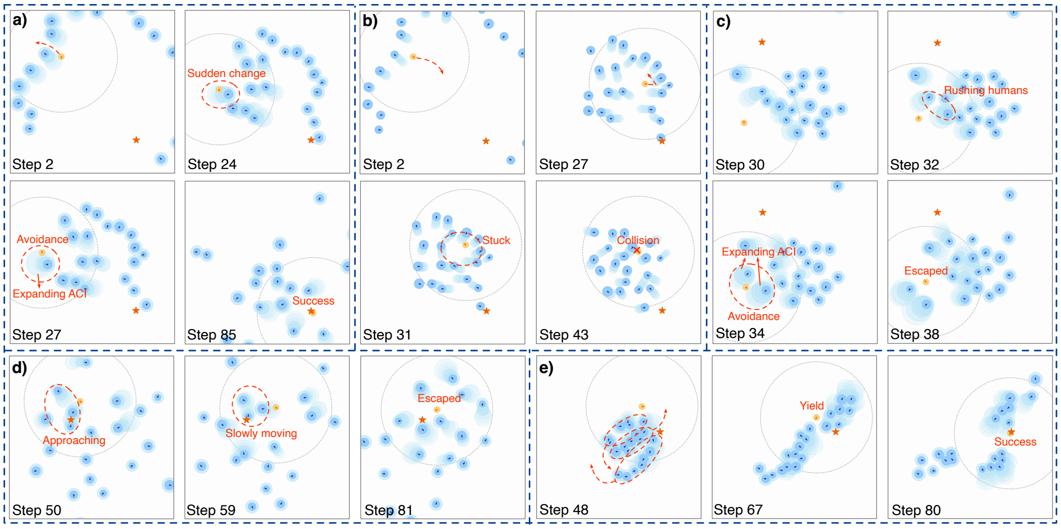

Figure 2: Test-case visualizations. Pedestrians are shown in blue, the robot in yellow, and the goal as an orange star. Light-blue circles show the quantified uncertainties around pedestrians. (a) Ours (w/ GST) successfully navigates to the goal in an in-distribution environment. (b) CrowdNav++ fails to complete the same episode; here, light-blue circles represent predicted trajectories rather than uncertainties. (c) Ours (w/ GST) in an OOD environment with rushing pedestrians. (d) Ours (w/ GST) in an OOD environment using the SF pedestrian model. (e) Ours (w/ GST) in an OOD environment with pedestrian groups.

For MPC, we implement a cost function similar to that in [43, 44, 45], which uses uncertainty as a reference in the cost formulation to improve safety. We test our method with the CV and the GST predictors, represented as Ours (w/ CV) and Ours (w/ GST), respectively. We also include RL (w/o ACI) and RL (w/ ACI) as two ablation models. The two variants have the same network structures as Ours (w/ GST), but RL (w/o ACI) excludes uncertainty estimates from the input and does not employ CRL, while RL (w/ ACI) incorporates uncertainty estimates but does not apply CRL.

4.3 In-Distribution Test Results

Quantitative Analysis. For all the test results in Table 1-2, we evaluate 1250 samples across 5 random test seeds and calculate the mean performance and standard deviations of models trained with 3 different training seeds. The test results under the same setting as the training environment are shown in Table 1. From these results, we can see that SF, ORCA, and MPC have lower SR and higher CR compared to RL methods. When comparing all the RL-based methods, both Ours (w/ GST) and Ours (w/ CV) outperform other methods in safety metrics and ITR, indicating that our method can generate both safe and polite trajectories that cause minimal intrusions to pedestrians. Although our method generates paths with higher TR, NT, and PL compared to RL baselines and ablation models, this tradeoff is reasonable considering the overall improvement in effectiveness and safety. Notably, Ours (w/ GST) achieves over an 8.80% higher SR, decreases collisions by more than 3.72 times, and reduces intrusions into the future pedestrian trajectories by more than 2.43 times compared to the two RL baselines.

Qualitative Analysis. We visualize the behaviors of Ours (w/ GST) and CrowdNav++ in the same episode in Fig. 2(a) and Fig. 2(b), respectively. At the beginning, CrowdNav++ decides to approach the goal directly. However, as the pedestrians move, they gradually surround the robot, which leads to an almost inevitable collision for CrowdNav++. In contrast, Ours (w/ GST) chooses to move the robot out of the crowds from the start. At step 24, Ours (w/ GST) rapidly reacts to a sudden change in human direction, and the adaptive expanding uncertainty area due to prediction errors helps the robot perform an avoidance maneuver. Additional qualitative examples are provided in our supplementary video.

4.4 Out-of-Distribution Test Results

OOD Scenarios Mixed with Rushing Humans. In this setting, we set 20% of the human agents to have a maximum speed of $2.0\,\mathrm{m/s}$. From the results in Table 2, we observe that all methods exhibit degraded performance compared to in-distribution conditions. Especially, SafeCrowdNav, CrowdNav++, RL (w/o ACI), and RL (w/ ACI) experience SR drops of 19.38%, 12.94%, 18.48%, and 17.12%, respectively. In contrast, Ours (w/ CV) and Ours (w/ GST) exhibit smaller SR drops of 8.96% and 9.76%, respectively. We visualize a scenario performed by Ours (w/ GST) in Fig. 2(c), where two rushing humans are moving toward the agent at step 32. Due to the GST predictor facing the challenges of OOD conditions, the uncertainty area generated by ACI becomes much larger. The robot then chooses to navigate between these two agents through the gap between the uncertainty areas at step 34 and successfully escapes to a safe area at step 38.

OOD Scenarios with SF Pedestrian Model. In this setting, we change the behavior policy of all human agents from ORCA to SF. From the results in Table 2, all methods perform better than in in-distribution conditions. Ours (w/ CV) and Ours (w/ GST) achieve almost perfect results in terms of SR, CR, and TR, implying that our methods *adapt well* to OOD scenarios caused by different behavior models. We visualize a scenario performed by Ours (w/ GST) in Fig. 2(d). When three humans approach each other, Ours (w/ GST) maintains a conservative distance from the crowds and successfully escapes afterward.

OOD Scenarios with Group Dynamics. In this setting, pedestrians are clustered into cohesive groups that maintain tight intra-group spacing. As Table 2 shows, every RL baseline and ablation variant suffers a noticeable performance drop. Especially, SafeCrowdNav and RL (w/ ACI) fall to SR below 76%. The adaptability to grouped pedestrians of CrowdNav++ is highly sensitive to the training seed, with variance exceeding 15%. By contrast, Ours (w/ CV) and Ours (w/ GST) retain SR above 94% and CR below 6%, while also displaying the smallest variance across seeds.

4.5 Other Test Results

We include additional test results in Appendix G to J, covering the effectiveness of ACI quantification, a convergence comparison between Ours (w/ GST) and RL (w/ ACI), the influence of different cost limits on policy aggressiveness, and cross-evaluations of algorithms trained in environments where humans either actively avoid or ignore the robot.

4.6 Real-Robot Experiments

We deploy our method on a ROSMASTER X3 robot with Mecanum wheels running ROS2, connected via router to a laptop equipped with a mobile NVIDIA RTX 3070 GPU. We apply the model to the robot without further fine-tuning, using only basic clipping and smoothing. Perception is performed with a 6 Hz RPLIDAR-A1 LiDAR, human detection via a pre-trained DR-SPAAM model [55], tracking with SORT [56], and trajectory prediction with GST. We implement two movement modes: fixed-goal reaching and long-range navigation with a dynamically updated goal. The robot navigates outdoor environments and executes safe behaviors in both sparse and dense crowds. For more details, please refer to Appendix K-M and the supplementary video.

5 CONCLUSION

In this paper, we introduce an RL-based trajectory-planning framework that integrates conformal uncertainty into a CRL scheme to mitigate OOD performance degradation. Unlike conventional RL planners that overfit and falter under distribution shifts, our method dynamically leverages uncertainty estimates to adapt to velocity variations, policy changes, and transitions from individual to group dynamics. Extensive simulations demonstrate robust stability across diverse OOD scenarios, and real-world trials confirm the practical effectiveness of our approach.

6 Limitations

First, our work addresses the safety challenges of navigating dense, human-populated environments by combining uncertainty-aware prediction with constrained reinforcement learning. It can also handle static obstacles when they are modeled as collections of static agents, as in prior work [11]. However, to naturally handle obstacles of arbitrary shape, our approach would require additional perception and mapping capabilities, such as integrating SLAM and feeding multi-modal observations into the policy network. This represents a clear direction for future research. The second limitation arises from perception errors (e.g., human miss-detections and the 2D LiDAR's sensitivity to lighting). To mitigate these issues, we plan to augment our sensor suite with cameras, apply sensor-fusion techniques, and upgrade to a higher-frequency, more robust LiDAR, thereby adding safety redundancy. Third, while our strategy effectively *mitigates* OOD performance degradation, it does not yet *eliminate* it, which is a great challenge in deep learning. Promising directions include adopting larger models (e.g., LLMs or VLMs) and advanced alignment methods, though these may introduce increased computational demands that could limit deployment on cost-effective robots. Finally, although we have evaluated OOD performance in three representative scenarios: velocity variations, policy changes, and transitions from individual to group-level dynamics, we acknowledge that the space of possible distributional shifts is vast and cannot be exhaustively tested. We plan to conduct larger-scale real-world experiments once static-obstacle handling is resolved. These real-world experiments represent the most challenging OOD setting. We will also employ metrics such as the intervention rate to further compare our method against standard RL baselines.

References

[1] A. Francis, C. Pérez-d'Arpino, C. Li, F. Xia, A. Alahi, R. Alami, A. Bera, A. Biswas, J. Biswas, R. Chandra, et al. Principles and guidelines for evaluating social robot navigation algorithms. *arXiv preprint arXiv:2306.16740*, 2023.

[2] 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*, 12(3):1–39, 2023.

[3] R. Zhou, H. Zhou, H. Gao, M. Tomizuka, J. Li, and Z. Xu. Grouptron: Dynamic multi-scale graph convolutional networks for group-aware dense crowd trajectory forecasting. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 805–811. IEEE, 2022.

[4] Z. Xu, R. Zhou, Y. Yin, H. Gao, M. Tomizuka, and J. Li. Matrix: multi-agent trajectory generation with diverse contexts. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12650–12657. IEEE, 2024.

[5] J. Van den Berg, M. Lin, and D. Manocha. Reciprocal velocity obstacles for real-time multi-agent navigation. In *2008 IEEE international conference on robotics and automation*, pages 1928–1935. Ieee, 2008.

[6] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51 (5):4282, 1995.

[7] J. R. Han, H. Thomas, J. Zhang, N. Rhinehart, and T. D. Barfoot. Dr-mpc: Deep residual model predictive control for real-world social navigation. *IEEE Robotics and Automation Letters*, 2025.

[8] S. Samavi, J. R. Han, F. Shkurti, and A. P. Schoellig. Sicnav: Safe and interactive crowd navigation using model predictive control and bilevel optimization. *IEEE Transactions on Robotics*, 2024.

[9] W. Jansma, E. Trevisan, Á. Serra-Gómez, and J. Alonso-Mora. Interaction-aware sampling-based mpc with learned local goal predictions. In *2023 International Symposium on Multi-Robot and Multi-Agent Systems (MRS)*, pages 15–21. IEEE, 2023.

[10] K. Li, Y. Chen, M. Shan, J. Li, S. Worrall, and E. Nebot. Game theory-based simultaneous prediction and planning for autonomous vehicle navigation in crowded environments. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2977–2984. IEEE, 2023.

[11] Y. F. Chen, M. Liu, M. Everett, and J. P. How. Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 285–292. IEEE, 2017.

[12] X. Ma, J. Li, M. J. Kochenderfer, D. Isele, and K. Fujimura. Reinforcement learning for autonomous driving with latent state inference and spatial-temporal relationships. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6064–6071. IEEE, 2021.

[13] S. Liu, P. Chang, W. Liang, N. Chakraborty, and K. Driggs-Campbell. Decentralized structural-rnn for robot crowd navigation with deep reinforcement learning. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 3517–3524. IEEE, 2021.

[14] H. Kim, K. Lee, J. Park, J. Li, and J. Park. Human implicit preference-based policy fine-tuning for multi-agent reinforcement learning in usv swarm. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

[15] Z. Xie and P. Dames. Drl-vo: Learning to navigate through crowded dynamic scenes using velocity obstacles. *IEEE Transactions on Robotics*, 39(4):2700–2719, 2023.

[16] Z. Xie and P. Dames. Scope: Stochastic cartographic occupancy prediction engine for uncertainty-aware dynamic navigation. *arXiv preprint arXiv:2407.00144*, 2024.

[17] K. Lee, J. Li, D. Isele, J. Park, K. Fujimura, and M. J. Kochendorfer. Robust driving policy learning with guided meta reinforcement learning. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 4114–4120. IEEE, 2023.

[18] M. Arief, M. Timmerman, J. Li, D. Isele, and M. J. Kochenderfer. Importance sampling-guided meta-training for intelligent agents in highly interactive environments. *IEEE Robotics and Automation Letters*, 2024.

[19] H. Ma, Y. Sun, J. Li, and M. Tomizuka. Multi-agent driving behavior prediction across different scenarios with self-supervised domain knowledge. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3122–3129. IEEE, 2021.

[20] J. Li, H. Ma, Z. Zhang, J. Li, and M. Tomizuka. Spatio-temporal graph dual-attention network for multi-agent prediction and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):10556–10569, 2021.

[21] S. Liu, P. Chang, Z. Huang, N. Chakraborty, K. Hong, W. Liang, D. L. McPherson, J. Geng, and K. Driggs-Campbell. Intention aware robot crowd navigation with attention-based interaction graph. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12015–12021. IEEE, 2023.

[22] H. Ma, Y. Sun, J. Li, M. Tomizuka, and C. Choi. Continual multi-agent interaction behavior prediction with conditional generative memory. *IEEE Robotics and Automation Letters*, 6(4): 8410–8417, 2021.

[23] J. Xu, W. Zhang, J. Cai, and H. Liu. Safecrowdnav: safety evaluation of robot crowd navigation in complex scenes. *Frontiers in neurorobotics*, 17:1276519, 2023.

[24] J. Li, D. Isele, K. Lee, J. Park, K. Fujimura, and M. J. Kochenderfer. Interactive autonomous navigation with internal state inference and interactivity estimation. *IEEE Transactions on Robotics*, 40:2932–2949, 2024.

[25] L. Wang, M.-A. Lavoie, S. Papais, B. Nisar, Y. Chen, W. Ding, B. Ivanovic, H. Shao, A. Abuduweili, E. Cook, et al. Deployable and generalizable motion prediction: Taxonomy, open challenges and future directions. *arXiv preprint arXiv:2505.09074*, 2025.

[26] Y. Wang, X. Huang, X. Sun, M. Yan, S. Xing, Z. Tu, and J. Li. Uniocc: A unified benchmark for occupancy forecasting and prediction in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2025.

[27] H. Girase, H. Gang, S. Malla, J. Li, A. Kanehara, K. Mangalam, and C. Choi. Loki: Long term and key intentions for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9803–9812, 2021.

[28] J. Li, X. Shi, F. Chen, J. Stroud, Z. Zhang, T. Lan, J. Mao, J. Kang, K. S. Refaat, W. Yang, et al. Pedestrian crossing action recognition and trajectory prediction with 3d human keypoints. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1463–1470. IEEE, 2023.

[29] J. Li, J. Li, S. Bae, and D. Isele. Adaptive prediction ensemble: Improving out-of-distribution generalization of motion forecasting. *IEEE Robotics and Automation Letters*, 2024.

[30] I. Gibbs and E. Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

[31] I. Gibbs and E. J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

[32] A. N. Angelopoulos, S. Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

[33] F.-Y. Sun, I. Kauvar, R. Zhang, J. Li, M. J. Kochenderfer, J. Wu, and N. Haber. Interaction modeling with multiplex attention. *Advances in Neural Information Processing Systems*, 35: 20038–20050, 2022.

[34] J. Li, F. Yang, M. Tomizuka, and C. Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Advances in neural information processing systems*, 33, 2020.

[35] V. M. Dax, J. Li, E. Sachdeva, N. Agarwal, and M. J. Kochenderfer. Disentangled neural relational inference for interpretable motion prediction. *IEEE Robotics and Automation Letters*, 9 (2):1452–1459, 2023.

[36] B. Lange, J. Li, and M. J. Kochenderfer. Scene informer: Anchor-based occlusion inference and trajectory prediction in partially observable environments. In *IEEE International Conferences on Robotics and Automation (ICRA)*, 2024.

[37] Z. Wang, Y. Wang, Z. Wu, H. Ma, Z. Li, H. Qiu, and J. Li. Cmp: Cooperative motion prediction with multi-agent communication. *IEEE Robotics and Automation Letters*, 2025.

[38] L. Luo and M. Cai. Bridging deep reinforcement learning and motion planning for model-free navigation in cluttered environments. *arXiv preprint arXiv:2504.07283*, 2025.

[39] C. Chen, Y. Liu, S. Kreiss, and A. Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *2019 international conference on robotics and automation (ICRA)*, pages 6015–6022. IEEE, 2019.

[40] J. Li, C. Hua, H. Ma, J. Park, V. Dax, and M. J. Kochenderfer. Multi-agent dynamic relational reasoning for social robot navigation. *arXiv preprint arXiv:2401.12275*, 2024.

[41] H. Ma, J. Li, R. Hosseini, M. Tomizuka, and C. Choi. Multi-objective diverse human motion prediction with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8161–8171, 2022.

[42] S. Yang, G. J. Pappas, R. Mangharam, and L. Lindemann. Safe perception-based control under stochastic sensor uncertainty using conformal prediction. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 6072–6078. IEEE, 2023.

[43] A. Dixit, L. Lindemann, S. X. Wei, M. Cleaveland, G. J. Pappas, and J. W. Burdick. Adaptive conformal prediction for motion planning among dynamic agents. In *Learning for Dynamics and Control Conference*, pages 300–314. PMLR, 2023.

[44] Z. Huang, T. Ji, H. Zhang, F. C. Pouria, K. Driggs-Campbell, and R. Dong. Interaction-aware conformal prediction for crowd navigation. *arXiv preprint arXiv:2502.06221*, 2025.

[45] L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas. Safe planning in dynamic environments using conformal prediction. *IEEE Robotics and Automation Letters*, 8(8):5116–5123, 2023.

[46] Y. Wang, S. Xing, C. Can, R. Li, H. Hua, K. Tian, Z. Mo, X. Gao, K. Wu, S. Zhou, et al. Generative ai for autonomous driving: Frontiers and opportunities. *arXiv preprint arXiv:2505.08854*, 2025.

[47] H. Huang, S. Sharma, A. Loquercio, A. Angelopoulos, K. Goldberg, and J. Malik. Conformal policy learning for sensorimotor control under distribution shifts. *arXiv preprint arXiv:2311.01457*, 2023.

[48] M. Golchoubian, M. Ghafurian, K. Dautenhahn, and N. L. Azad. Uncertainty-aware drl for autonomous vehicle crowd navigation in shared space. *IEEE Transactions on Intelligent Vehicles*, 2024.

[49] K. J. Strawn, N. Ayanian, and L. Lindemann. Conformal predictive safety filter for rl controllers in dynamic environments. *IEEE Robotics and Automation Letters*, 2023.

[50] C. Schöller, V. Aravantinos, F. Lay, and A. Knoll. What the constant velocity model can teach us about pedestrian motion prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, 2020.

[51] Z. Huang, R. Li, K. Shin, and K. Driggs-Campbell. Learning sparse interaction graphs of partially detected pedestrians for trajectory prediction. *IEEE Robotics and Automation Letters*, 7(2):1198–1205, 2021.

[52] A. Ray, J. Achiam, and D. Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 2019.

[53] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[54] J. Ji, J. Zhou, B. Zhang, J. Dai, X. Pan, R. Sun, W. Huang, Y. Geng, M. Liu, and Y. Yang. Omnisafe: An infrastructure for accelerating safe reinforcement learning research. *arXiv preprint arXiv:2305.09304*, 2023.

[55] D. Jia, A. Hermans, and B. Leibe. Dr-spaam: A spatial-attention and auto-regressive model for person detection in 2d range data. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10270–10277. IEEE, 2020.

[56] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

Appendices for *Towards Generalizable Safety in Crowd Navigation via Conformal Uncertainty Handling*

A Preliminaries

Adaptive Conformal Inference. Conformal methods can augment model predictions with a prediction set that is guaranteed to contain true values with a predefined coverage, enabling the quantification of uncertainties in a model-agnostic manner [31]. Traditional split conformal prediction requires a calibration set and places high demands on the exchangeability between the test sample and the calibration samples. In contrast, adaptive conformal inference (ACI) can dynamically adjust its parameters to maintain coverage in an online and distribution-free manner [30], making it appealing for time-sequential applications. Dynamically-tuned adaptive conformal inference (DtACI) [31] further boosts the applicability and performance of ACI by running multiple prediction error estimators with different learning rates simultaneously. DtACI adaptively selects the best output based on its historical performance, eliminating the need to pre-acquire underlying data dynamics to achieve satisfying coverage.

Constrained Reinforcement Learning. Constrained reinforcement learning (CRL) extends RL algorithms by incorporating constraints on the agent's behavior. Unlike traditional Markov decision process (MDP) settings, where agents learn behaviors only to maximize rewards, CRL is often formulated as a constrained Markov decision process (CMDP). At time step t, an agent chooses an action A_t under state S_t, receives a reward R_t, and incurs a cost C_t, after which the environment transitions to the next state S_{t+1}. In a CMDP, the objective is not only to find an optimal policy that maximizes rewards but also to manage costs associated with certain actions or states, which may be defined as quantities related to safety in the context of social navigation. This is generally represented as [52]:

$$\pi^* = \arg\max_{\pi \in \Pi^C} J^R(\pi), \tag{12}$$

where $J^R(\pi)$ is a reward-based objective function, and Π^C is the feasible set of policies that satisfy the constraints added to the problem. The goal of CRL is to ensure that costs remain within predefined thresholds while maximizing reward.

B Conceptual Differences with CBF

In CBFs, a safety function $h(x)$ and a safe set $\{x : h(x) > 0\}$ are used to encode safety. The condition $\dot{h}(x) \geq -\gamma h(x)$ is enforced during optimization to ensure the system remains in the safe set, constraining the *derivative* of the safety function. In contrast, our framework focuses on controlling the *integral* of a risk-related function. Our cost function $C(S, A)$, analogous to $h(x)$, is accumulated over time as $\sum_{t=0}^{T} C_t(S_t, A_t)$, and we aim to constrain this cumulative cost below a small threshold $\tilde{d}_C$. Thus, while both approaches can shape the distribution of a safety-related function, the optimization mechanisms differ significantly. Our methods integrate more closely with CRL and achieve expected safety.

C SF vs. ORCA

We plotted human trajectories generated using ORCA and SF dynamics in Fig. 3, along with plots of minimal distances between individuals and mean velocities. Note that in our simulation environments, humans may sample a new goal every five steps with a probability of 50%, which changes their moving directions and velocities. From the results, ORCA pedestrians exhibit sharper turns and smaller minimal distances between individuals.

D Evaluation Metrics

The evaluation metrics used in our paper are introduced in detail as follows:

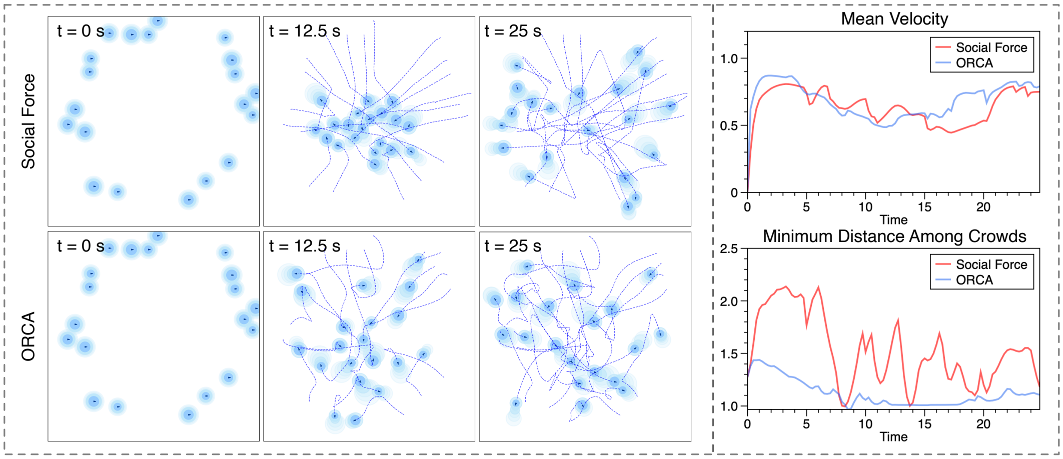

Figure 3: Comparison of SF and ORCA.

- *Success Rate (SR):* SR measures the ratio of the number of successful episodes (i.e., the robot reaching the goal within the time limit of $50\,\text{s}$) to the total number of test episodes, i.e., SR $= N_{\text{success}}/N_{\text{total}}$.
- *Collision Rate (CR:* CR measures the ratio of episodes with at least one collision to the total number of test episodes, i.e., CR $= N_{\text{collision}}/N_{\text{total}}$.
- *Timeout Rate (TR):* TR measures the ratio of episodes that fail to reach the goal before the time limit to the total number of test episodes, i.e., TR $= N_{\text{timeout}}/N_{\text{total}}$.
- *Navigation Time (NT):* NT is the average time required for the robot to reach the goal (computed only over *successful* episodes), i.e., NT $= \big(1/N_{\text{success}}\big)\sum_{k=1}^{N_{\text{success}}} T_k$, where T_k is the time the robot took to reach the goal in the k-th successful episode.
- *Path Length (PL):* PL is the total distance traveled by the robot in an episode, accumulated step by step. We report the average path length across all episodes (including collisions and timeouts), i.e., PL $= \big(1/N_{\text{total}}\big)\sum_{i=1}^{N_{\text{total}}} \text{dist}_i$, where dist_i is the total distance traveled in episode i.
- *Intrusion Time Ratio (ITR):* ITR is the fraction of time steps in an episode where the robot intrudes into the ground-truth future positions of humans (*Danger* event triggered). Its final value is an average over all episodes:

$$\text{ITR} = \frac{1}{N_{\text{total}}}\sum_{i=1}^{N_{\text{total}}}\Big(\frac{\#\text{ danger steps in episode } i}{\text{total steps in episode } i} \times 100\%\Big). \tag{13}$$

- *Social Distance (SD):* SD is the average of the minimal distances between the robot and any human, computed only at the "danger steps." We log the minimal distance whenever a *Danger* event occurs, and then take the average over all episodes.

E Implementation Details

Under each ablation setting, we train our models using three different random seeds on an NVIDIA RTX 4090 GPU, keeping all other hyperparameters consistent except those we intend to compare. Key common settings include:

- All models use human predictions as part of the input observations and employ a pre-trained GST predictor with fixed parameters to generate five steps of human trajectory predictions, except for Ours (w/ CV).

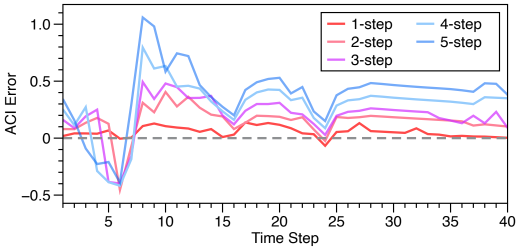

Figure 4: Visualization of ACI errors (i.e., estimated prediction error minus actual prediction error) for one pedestrian's five prediction steps during the time period it is within the observable area of the robot. ACI provides valid coverage when the ACI error is greater than 0.

- For the training parameters, we leverage the parallel training feature of PPO by and running 128 parallel environments. The learning rate is set to 3×10^{-5}, and the clip parameter is set to 0.08 for both the actor and the reward critic. The learning rate for the cost critic is set to 1.5×10^{-5}. We run 2×10^7 training steps to obtain the final policy model for testing and real-robot deployment. In our training, varying entropy is removed, and the action noise, which is normally distributed, is only added during training and removed during testing and deployment.
- We set the radius of the cost subareas around current positions of humans, r_{comfort}, to $0.25\,\text{m}$ and calculate costs based on the intrusions into the subarea and the first two prediction uncertainty areas, while inputting predictions for all five future steps into the policy network with a coverage parameter $\alpha = 0.1$, corresponding to 90% coverage in prediction errors.
- For DtACI hyperparameters, the initial prediction errors are set to $0.1\,\text{m}$, $0.2\,\text{m}$, $0.3\,\text{m}$, $0.4\,\text{m}$, and $0.5\,\text{m}$ for 1- to 5-step-ahead predictions, respectively. We employ three error estimators with learning rates γ of 0.05, 0.1, and 0.2 for each DtACI estimator.
- For the specific reward and cost functions, we set the success reward R_{success} to $+10$, the collision penalty $R_{\text{collision}}$ to -20, and the dense potential reward that guides the robot's movement to $2\Delta d_{\text{forward}}$, where $\Delta d_{\text{forward}}$ measures the distance the robot moves toward the goal compared to the previous step. The cost $C_t(S_t, A_t)$ is defined as $2.5 d_{\text{intru},t}$, where $d_{\text{intru},t}$ represents the maximum intrusion into the cost areas of the current human positions and the first two prediction uncertainty areas.

F Tradeoffs of Different Methods

From the results in Tables 1-2, we observe that: (1) SF, ORCA, and MPC show poor performance across multiple metrics, particularly in collision avoidance and goal-reaching capabilities. (2) CrowdNav++, RL (w/o ACI), and RL (w/ ACI) achieve reasonable goal-reaching rates and safety performance. (3) SafeCrowdNav achieves comparable task completion and safety performance with shorter path lengths, but exhibits higher intrusion rates, indicating less socially-aware behavior. (4) Our method achieves the highest goal-reaching rates with the lowest collision rates and fewest intrusions, at the cost of longer paths and navigation time, since our framework prioritizes safety over efficiency. However, this safety-efficiency tradeoff can be easily adjusted via the cost limit parameter for different deployment scenarios.

G ACI Effectiveness

To validate the effectiveness of ACI in covering actual prediction errors, we visualize the ACI errors of one human in Fig. 4. At the beginning of this trajectory, the ACI errors for multi-step predictions are large because the GST predictor lacks sufficient information to accurately predict human positions, leading to large actual prediction errors that our initial ACI values do not cover properly.

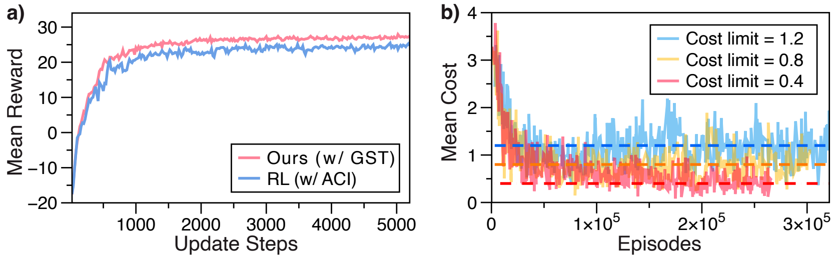

Figure 5: Convergence analysis of our method. (a) The learning curves of Ours (w/ GST) and RL (w/ACI). Ours (w/ GST) shows faster convergence with higher rewards. (b) The cost curves of Ours (w/ GST) with different cost limits. The average costs across episodes can approximately approach the predefined cost limits, which are shown by the dashed lines.

Table 3: Performance of Ours (w/ GST) Under Different Cost Limits

Cost limit	**SR↑**	**CR↓**	**TR↓**	**NT↓**	**PL↓**	**ITR↓**	**SD↑**
0.4	96.96%	2.80%	0.24%	18.52	25.24	2.73%	0.43
0.6	96.72%	3.04%	0.24%	16.16	22.96	3.59%	0.44
0.8	94.56%	5.44%	0.00%	13.60	20.33	5.61%	0.43
1.0	94.32%	5.68%	0.00%	13.53	20.19	6.62%	0.42
1.2	93.76%	6.24%	0.00%	12.97	19.45	7.31%	0.41

However, after several steps, ACI quickly adapts to the actual prediction error and achieves adequate coverage. Additionally, if the coverage remains sufficient, the ACI error will decrease to ensure that the uncertainty estimation is not too conservative. We also extract human trajectories and the corresponding ACI predictions to evaluate the quantitative coverage performance. The coverage rates for 1-, 2-, 3-, 4-, and 5-step predictions are 90.21%, 91.53%, 92.34%, 90.78%, and 90.55%, respectively, using the parameters specified in our approach with $\alpha = 0.1$, which theoretically corresponds to 90% coverage.

H Convergence Analysis

We present the learning curves of Ours (w/ GST) in Fig. 5(a), alongside RL (w/ ACI) for comparison. The results indicate that Ours (w/ GST) demonstrates a smoother learning process and achieves higher rewards. Furthermore, as shown in Fig. 5(b), the average episodic costs of Ours (w/ GST) converge to different cost limit values. This provides an effective mechanism for adjusting the aggressiveness of robot policies, as further discussed in Appendix I.

I Tuning Aggressiveness of Trajectories with Cost Limits

Although the main results presented in our paper focus on low cost limits to ensure safe navigation through crowds, we also observe interesting testing results with varying cost limits that allow for tuning the aggressiveness of trajectories, as shown in Table 3. These results are obtained by testing across five seeds, with 250 test samples for each seed, using models trained with the same training seed. We find that as the cost limits increase, the trajectories become gradually more aggressive, reflected by an increase in ITR. This increased aggressiveness also improves efficiency, as indicated by reductions in NT and PL. However, it comes at the expense of key metrics such as SR and CR. Therefore, for a single best policy, we recommend keeping the cost limit as low as possible, provided the policy can still converge.

J Visible Robot Settings

For the main results presented in our paper, we train the RL baselines, ablation models, and our methods in the CrowdNav environments where humans will not actively avoid the robot, which is

Table 4: Cross-Validation of Ours (w/ GST) for Robot Visibility Settings Using Success Rates

Train Test	Invisible	Visible
Invisible	**96.93±0.68%**	50.48±0.49%
Visible	**99.92±0.14%**	99.07±0.76%

Table 5: Cross-Validation of Ours (w/ GST) for Robot Visibility Settings Using Collision Rate

Train Test	Invisible	Visible
Invisible	**2.93±0.61%**	49.52±0.49%
Visible	**0.08±0.14%**	0.93±0.76%

a common practice in prior work [21, 23]. To validate the effectiveness of this setting, we conduct a cross-evaluation on models trained in environments where humans either actively avoid or ignore the robot, as summarized in Tables 4-5. Specifically, we evaluate the models under five different testing seeds, each with 250 test samples, for three models trained under three distinct training seeds. From these results, we observe that although models trained in the visible robot setting achieve low collision rates in their in-distribution settings, they fail to generalize when humans do not react to the robot. In such scenarios, nearly half of the test cases result in collisions. In contrast, models trained in the invisible robot setting maintain low collision rates in both testing scenarios and, surprisingly, exhibit even lower collision rates when the robot is visible than those trained in the visible robot setting. In real-world situations, some pedestrians may ignore the robot (e.g., when they are using their mobile phones while walking). This mixture of reactive and non-reactive pedestrians poses significant challenges for models that assume all humans will actively avoid the robot. Thus, training in an environment where humans do not react to the robot proves to be more robust, as it better accounts for scenarios in which pedestrians may fail to notice the robot.

K ROS2 System for Real Robot Deployment

We develop a full ROS2 system from perception to decision making, including four main nodes:

- *Detector:* This node employs a pretrained DR-SPAAM model [55] to detect human agents from 2D LiDAR point clouds. The model leverages a lightweight neural network, enabling real-time detection on resource-constrained devices. According to the original paper, it achieves an accuracy metric of $AP_{0.5} = 70.3\%$.
- *Tracker:* This node uses the SORT [56] algorithm for tracking, which simply aims at assigning indices to detected agents. To generate consistent tracking results even when receiving noisy detections, we assign large values to the measurement noise covariance matrix $\mathbf{R}$ and the process noise covariance matrix $\mathbf{Q}$ of the underlying Kalman filter to accommodate detection jitter. We also set a higher noise coefficient for the position and scale components of the Kalman filter and a smaller value for the velocity components to allow more flexibility in position and scale updates. In the initial state covariance matrix $\mathbf{P}$, the velocity components are given a large uncertainty, and the entire matrix is scaled accordingly so the filter can quickly adapt during early tracking. Besides, in the SORT tracker, we only keep trajectories whose ratio of valid detections over the entire detected period is above 0.9, filtering out spurious or short-lived detections. At each timestep, we run the Kalman filter's inference step to infer the positions of tracked agents even when they are not correctly detected in certain frames.
- *Predictor:* This node uses the GST [51] model as the trajectory planner, integrating the DtACI module to directly apply uncertainty quantification after obtaining prediction lines for the pedestrians. The parameters and model weights are consistent with those used during training.

- *Decider:* This node receives information from the upper-level nodes and determines the output action commands (v_x and v_y for velocities) for the controller. We utilize the controller integrated into the ROSMASTER X3 system by publishing `\cmd_vel`. The decider supports three modes: goal-reaching mode, long-range navigation mode, and manual control mode. The selection of these modes is managed by another utility node, *Command Listener*, which listens to user input and communicates the commands to the decider.

L Real Robot Experiments

We deploy our approach on a ROSMASTER X3 robot equipped with Mecanum wheels, enabling flexible movement and independent control of v_x and v_y. The robot connects to a laptop with an Nvidia RTX 3070 GPU (mobile version) via a router. Our model is trained in CrowdNav simulations with holonomic dynamics and is deployed directly to the robot system *without* further fine-tuning.

The robot uses a 2D LiDAR (RPLIDAR-A1) for human detection via a learning-based detector. The LiDAR operates at a scanning frequency of approximately $6\,\mathrm{Hz}$, which also limits the frequencies of the tracking, prediction, and decision making modules. Despite this, our experiments demonstrate that the robot makes effective navigation decisions in crowded environments, indicating that the system generalizes well to cost-efficient robots. Human detection is done with a pre-trained DR-SPAAM model [55], tracking is done with SORT [56], prediction is done with the GST model [51], and decision making is powered by Ours (w/GST). For further details about our system, please refer to Appendix K, and computational speed analyses for the core functions of each module are provided in Appendix M.

We conduct the experiments on a large outdoor terrace covering an area greater than $15\,\mathrm{m} \times 20\,\mathrm{m}$. The robot operates in two movement modes:

- *Goal-reaching mode:* The robot navigates to a predetermined target point, which is the original setting in the standard CrowdNav training phase.
- *Long-range navigation mode:* The goal of the robot is dynamically updated based on its movement. This approach addresses the limitation that, during training, the goal inputs to the policy networks are constrained to a limited range. As a result, distant goals beyond $12\,\mathrm{m}$ may negatively impact policy performance. By introducing a continuously updated moving target, we show that the robot can reliably navigate distances exceeding $20\,\mathrm{m}$ through experiments. This method has the potential to be extended to even longer distances if the system is fully integrated onboard in the future.

We show several representative real-world testing scenarios and results in Fig. 6. Specifically, we first visualize the prediction uncertainties using RViz in Fig. 6(a). Before $t = 13\,\mathrm{s}$, the human walks around the robot, and the robot successfully detects the human's position while visualizing the predicted trajectory and the associated uncertainty area. Since the prediction model does not fully capture the human's circular walking pattern, the long-horizon prediction uncertainty area is significantly larger than that of short-horizon predictions. After $t = 13\,\mathrm{s}$, the human stands still behind the robot. Initially, the prediction uncertainty area is large because the uncertainty results are carried over from iterative processes. Gradually, the uncertainty area shrinks to reflect more precise prediction uncertainties. After approximately 3 seconds, the uncertainty area stabilizes and reduces to almost zero. This process demonstrates that by using DtACI, the system can effectively adapt to different conditions and dynamically reflect prediction uncertainties. This enables safer decision making without being overly conservative.

In Fig. 6(b), by allowing the robot to interact with a single agent, we observe that the robot can generate stable yielding behaviors. At $t = 2\,\mathrm{s}$, the robot detects that a human is coming along a path that will intersect with its original moving trajectory. It begins by slowing down to reduce the risk of a collision ($t = 3\,\mathrm{s}$). At $t = 3.5\,\mathrm{s}$, the robot adjusts its direction, moving towards the backside area of the human to avoid intrusions. At $t = 4.5\,\mathrm{s}$, the robot successfully passes.

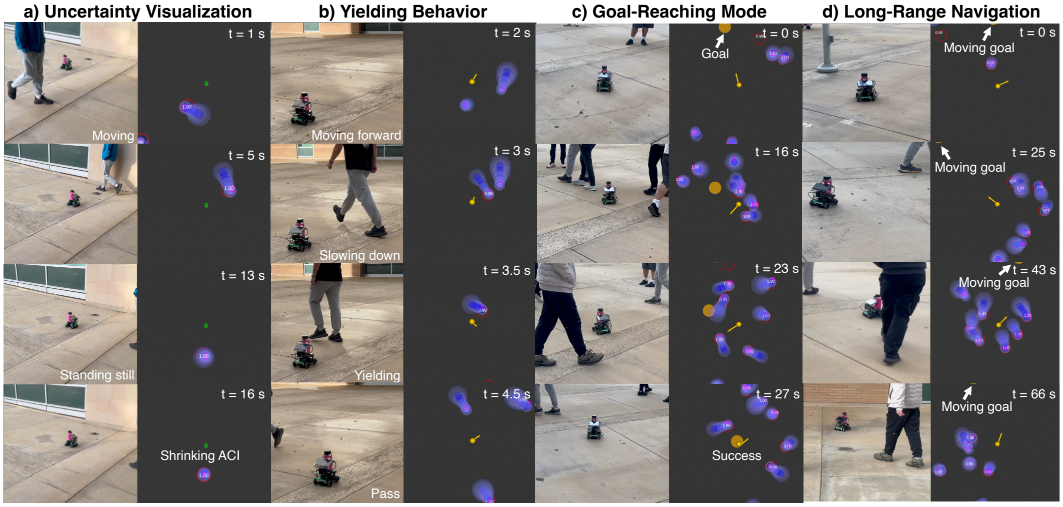

Figure 6: We deploy our methods on a ROSMASTER X3 with Mecanum wheels using the ROS2 system. For the four subplots, the left sides display photos taken from the experiments, and the right sides show visualizations in RViz. In the RViz visualizations, the red circles represent detection results, and the white numbers inside the red circles indicate the output probabilities of the detection model. The purple numbers correspond to the indices generated by the tracker. The prediction lines are shown in blue, and the prediction uncertainties are depicted in semi-transparent light blue. In (a), the green robot represents the robot's current location. In (b)-(d), where the decision node is enabled, the yellow sphere indicates the robot's position, and the yellow arrow represents the command output from the decider node. The orange circle indicates the goal position. (a) In the uncertainty visualization, the human initially walks around the robot and then stands still behind it. The uncertainty area adjusts dynamically based on the prediction accuracy. (b) The robot equipped with Ours (w/ GST) demonstrates stable yielding behavior when interacting with humans. (c) In goal-reaching mode, the robot navigates through crowds and successfully reaches its goal. (d) In long-range navigation mode, the moving goal consistently guides the robot's movement.

In Fig. 6(c), the robot navigates in goal-reaching mode while interacting with humans along its path. We observe that the robot encounters very dense crowds at $t = 16\,\mathrm{s}$ and $t = 23\,\mathrm{s}$. Throughout this process, the robot makes appropriate decisions to minimize intrusions into humans' trajectories while progressing toward its goal, and it ultimately reaches the goal despite the dense human interactions.

In Fig. 6(d), we implement the long-range navigation mode, enabling the robot to navigate distances beyond those encountered during training. In this mode, the moving goal is dynamically updated to remain $5\,\mathrm{m}$ ahead of the robot's position along the longitudinal axis, while its lateral position remains fixed. Under these conditions, the robot moves forward while actively avoiding dynamic obstacles, as demonstrated at $t = 25\,\mathrm{s}$. Although the robot temporarily deviates laterally to avoid collisions, the fixed lateral component of the moving goal serves as a stable reference, helping the robot stay within a reasonable lateral range.

Overall, our system demonstrates robust decision making capabilities in crowded environments. By quantifying prediction uncertainty and incorporating it into behavior guidance, the robot develops more adaptive spatial awareness and exhibits consistent, safe, and socially aware behaviors.

M Computational Speed Analysis

We log the time consumption of key functions across different nodes, including the inference times of the detection model (DR-SPAAM), prediction model (GST), and the policy network (CRL), as well as the update times for the tracker (SORT) and uncertainty quantifier (DtACI) in dense crowd experiments, and visualize the results in Fig. 7. The visualization shows that GST has the longest

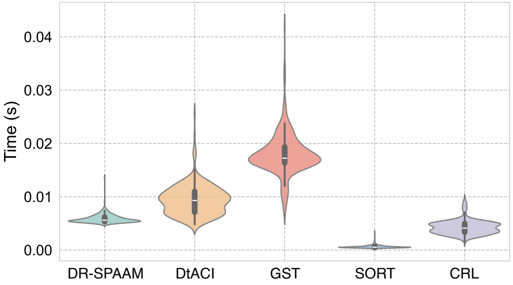

Figure 7: Visualizations of computational time distributions for five key modules of our ROS2 system. The data samples were collected during experiments in dense human environments. The shape of each violin plot represents the density of computation times, with wider areas indicating a higher concentration of values. The white dots at the center of each shape denote the mean computation time for the respective module.

average computation time and the largest variance. In contrast, DR-SPAAM and SORT maintain low and stable computation times, making them efficient for real-time applications. While DtACI has relatively high computational costs due to its pure Python implementation, its average time remains below $0.01\,\mathrm{s}$, well within acceptable limits. This overhead can be greatly reduced by transitioning to more efficient languages like C++. Despite this, the worst-case computation time for all main modules stays below 0.1 seconds, meeting and exceeding our system's decision frequency requirements.