2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Learning Social Navigation from Positive and Negative Demonstrations and Rule-Based Specifications

Chanwoo Kim[1], Jihwan Yoon[1], Hyeonseong Kim[1], Taemoon Jeong[1], Changwoo Yoo[2], Seungbeen Lee[3,4], Soohwan Byeon[5], Hoon Chung[5], Matthew Pan[6], Jean Oh[7], Kyungjae Lee[8], and Sungjoon Choi[1*]

***Abstract*— Mobile robot navigation in dynamic human environments requires policies that balance adaptability to diverse behaviors with compliance to safety constraints. We hypothesize that integrating data-driven rewards with rule-based objectives enables navigation policies to achieve a more effective balance of adaptability and safety. To this end, we develop a framework that learns a density-based reward from positive and negative demonstrations and augments it with rule-based objectives for obstacle avoidance and goal reaching. A sampling-based lookahead controller produces supervisory actions that are both safe and adaptive, which are subsequently distilled into a compact student policy suitable for real-time operation with uncertainty estimates. Experiments in synthetic and elevator co-boarding simulations show consistent gains in success rate and time efficiency over baselines, and real-world demonstrations with human participants confirm the practicality of deployment. A video illustrating this work can be found on our project page `https://chanwookim971024.github.io/PioneeR/`.**

I. Introduction

Mobile robot navigation in crowded, human-shared environments is inherently safety-critical and requires policies that remain reliable while adapting to diverse human behaviors. Core challenges [1], [2] include uncertainty in human intent, variability in motion patterns, dense interactions and bottlenecks, compliance with social conventions and right-of-way, and strict real-time requirements on embedded platforms. Addressing these challenges is essential for socially aware and reliable navigation.

Classical approaches [3]–[8] provide interpretability and explicit safety guarantees but often rely on carefully specified objectives and handcrafted rules, making them difficult to generalize in socially dynamic contexts [1]. Learning-based methods [9]–[12] instead seek to capture human interaction patterns directly from data, enabling adaptive and socially responsive behaviors. However, reinforcement learning typically demands extensive reward shaping and large training budgets, while imitation learning is more data-efficient [13] yet remains prone to distributional shifts and lacks explicit safety mechanisms [14], [15]. These limitations motivate designs that combine the adaptability of learning with the reliability of explicit safety specifications.

We hypothesize that integrating data-driven rewards with rule-based objectives enables navigation policies that achieve a more effective balance of adaptability and safety. To this end, we develop a framework that learns a density-based reward map from positive and negative demonstrations and augments it with rule-based objectives for obstacle avoidance and goal reaching. A teacher policy evaluates short-horizon rollouts under this formulation, producing supervisory actions that are both adaptive to demonstrated behaviors and explicitly safe by design. For real-time deployment, the teacher is distilled into a compact student policy that conditions directly on observations, inheriting adaptability and safety while remaining suitable for embedded operation.

The main contributions of this work are threefold: (i) a unified reward formulation that integrates positive and negative demonstration-driven density learning with rule-based safety specifications for obstacle avoidance and goal reaching; (ii) a teacher policy built on this formulation that provides adaptive and safe supervision, together with an uncertainty-aware distillation process that yields a compact policy for real-time operation; and (iii) evaluation in elevator co-boarding scenarios, in both simulation and real-world trials, assessing the effectiveness of combining data-driven rewards with rule-based safety in dynamic human environments.

II. Related Work

A. Navigation in Socially Dynamic Environments

Navigation in socially dynamic environments has been studied using classical, learning-based, and hybrid approaches. Classical methods, including window-based search [3], potential fields [4], velocity–obstacle formulations [5], and control barrier functions [6]–[8], provide interpretable objectives and explicit safety constraints, but rely on handcrafted cost designs that can be difficult to scale to complex human interactions [1]. Learning-based approaches instead infer interaction patterns directly from data [9]–[12]. Reinforcement learning can produce socially compliant behaviors but often requires substantial reward engineering

[1]Chanwoo Kim, Jihwan Yoon, Hyeonseong Kim, Taemoon Jeong, and Sungjoon Choi are with the Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea. (e-mails: {chanwoo-kim, yoonmungchi, hyeonseong-kim, taemoon-jeong, sungjoon-choi}@korea.ac.kr)

[2]Changwoo Yoo is with the Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea. (e-mail: cwyoo01@korea.ac.kr)

[3]Seungbeen Lee is with the Department of Artificial Intelligence, Yonsei University, Seoul, Republic of Korea. (email: seungblee@yonsei.ac.kr)

[4]Seungbeen Lee is with the Robotics Institute, School of Computer Science at Carnegie Mellon University, Pittsburgh, PA, USA. (email: seungbel@andrew.cmu.edu)

[5]Soohwan Byeon and Hoon Chung are with Mobinn, Suwon, Republic of Korea. (e-mail: {soohwan.byeon, h.chung}@mobinn.co.kr)

[6]Matthew Pan is with the Department of Electrical and Computer Engineering, Queen's University, Kingston, Canada. (e-mail: matthew.pan@queensu.ca)

[7]Jean Oh is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA 15213. (e-mail: jeanoh@cs.cmu.edu)

[8]Kyungjae Lee is with the Department of Statistics, Korea University, Seoul, Republic of Korea. (email: kyungjae_lee@korea.ac.kr)

and training effort, while imitation learning is more data-efficient [13] yet remains sensitive to distributional shifts and lacks explicit safety structure [14], [15]. Hybrid frameworks attempt to combine learning with rule-based or optimization modules [16]–[19], improving robustness but still relying on manually specified logic or simplified evaluation settings.

B. Data-Driven Reward Learning

Another line of work focuses on learning reward representations from demonstrations of varying quality. Smooth leveraged kernels [20], [21] incorporate demonstration quality to improve robustness, while suboptimal or unsafe trajectories can provide counterexamples that delineate undesirable behaviors [20]–[22]. Density-matching reward learning [23] aligns state–action visitation distributions, producing occupancy-like reward structures that emphasize feasible behaviors. Building on these ideas, our approach constructs a density-based reward from positive and negative demonstrations and integrates it with rule-based terms for obstacle avoidance and goal reaching, enabling navigation that balances adaptability and safety in dynamic human environments.

III. Problem Formulation

A. State and Observation

Let $x \in \mathbb{R}^n$ denote the robot state and $u \in \mathcal{U} \subset \mathbb{R}^m$ the control input. The robot dynamics are described by $\dot{x} = f(x, u)$, and at each decision step the robot issues a velocity command

$$u = [v, \omega]^\top \in \mathcal{U}, \qquad v \in \mathbb{R}_{\geq 0}, \;\; \omega \in \mathbb{R}, \tag{1}$$

where v and ω denote translational and rotational velocities. We assume a unicycle model with state $x = (p_x, p_y, \theta) \in \mathbb{R}^2 \times \mathbb{S}^1$.

The observation $o \in \Omega$ is constructed from LiDAR measurements and geometric descriptors. The LiDAR scan with G beams is grouped into $K = G/g$ segments, retaining the minimum range from each group to produce $\{r(\theta_1), \ldots, r(\theta_K)\}$. We additionally include the distances and relative angles of the b nearest obstacles $\{(d_i, \phi_i)\}_{i=1}^{b}$ and the relative goal angle ϕ_g, forming

$$o = \Big[r(\theta_1), \ldots, r(\theta_K),\; d_1, \ldots, d_b,\; \phi_1, \ldots, \phi_b,\; \phi_g\Big]. \tag{2}$$

In our implementation, $G = 72$, $g = 3$, and $b = 2$, resulting in $o \in \mathbb{R}^{29}$.

B. Problem Statement

We address the navigation problem of a mobile robot operating in dynamic human environments. The goal is to learn a parameterized policy

$$\pi_\theta : \Omega \to \mathcal{U}, \tag{3}$$

that maps observations $o \in \Omega$ to velocity commands $u \in \mathcal{U}$, enabling safe and adaptive closed-loop navigation. To this end, we adopt a teacher–student formulation: the teacher policy evaluates candidate actions using a reward constructed from positive and negative demonstrations together with rule-based specifications, while the student policy π_θ distills this supervision into a compact controller suitable for real-time deployment.

IV. Proposed Method

We present PioneeR (**P**ositive and negat**i**ve dem**o**nstration de**n**sity–driv**e**n **r**ewards with **R**ule-based specifications), a framework for mobile robot navigation in dynamic human environments. PioneeR constructs a reward map by combining density-based rewards inferred from positive and negative demonstrations with rule-based objectives for obstacle avoidance and goal progress. This representation encodes human-informed navigation preferences while enforcing safety requirements. A sampling-based lookahead controller evaluates candidate rollouts on this reward and selects actions with the highest return, as illustrated in Fig. 1.

The resulting teacher policy benefits from forward simulation to generate safe and adaptive supervision but relies on privileged information through forward simulation of future states. To address this, we distill the teacher into a student policy that conditions only on observations. The distilled policy retains the adaptability and safety of the teacher and additionally outputs uncertainty estimates that provide indicators of navigation risk in dynamic environments.

A. Teacher Policy via Reward Design

The teacher evaluates short-horizon rollouts generated from sampled velocity commands and selects the first action with the highest return. The return is computed from two components: (i) a reward learned from positive and negative demonstrations and (ii) rule-based specifications encoding safety and task objectives. This design produces trajectories that reflect demonstrated navigation patterns while maintaining explicit safety margins.

1) Density Reward Learning from Positive and Negative Demonstrations: We learn a reward over state-action pairs tailored to navigation by aligning reward values with the empirical occupancy of demonstrated behavior. Let $\mathcal{S}$ and $\mathcal{A}$ denote state and action spaces, and write $x = (s, a) \in \mathcal{X} := \mathcal{S} \times \mathcal{A}$. From demonstrations $\{x_j\}_{j=1}^{N_D} \subset \mathcal{X}$, define the empirical state–action density $\hat{\mu}$. The reward $R : \mathcal{X} \to \mathbb{R}$ is obtained by maximizing its expected value under $\hat{\mu}$,

$$\max_{R} \;\langle \hat{\mu}, R \rangle \quad \text{subject to} \quad \|R\|_2 \leq 1, \tag{4}$$

where

$$\langle \hat{\mu}, R \rangle \;=\; \int_{\mathcal{S}\times\mathcal{A}} \hat{\mu}(x)\, R(x)\, dx. \tag{5}$$

This formulation is model-free, relying on demonstrations.

For practical computation and generalization, we represent R in a reproducing kernel Hilbert space (RKHS) with positive semidefinite kernel k. Using inducing points $U = \{u_i\}_{i=1}^{N_U} \subset \mathcal{X}$ and coefficients $\alpha \in \mathbb{R}^{N_U}$,

$$R(x) \;=\; \sum_{i=1}^{N_U} \alpha_i\, k(x, u_i). \tag{6}$$

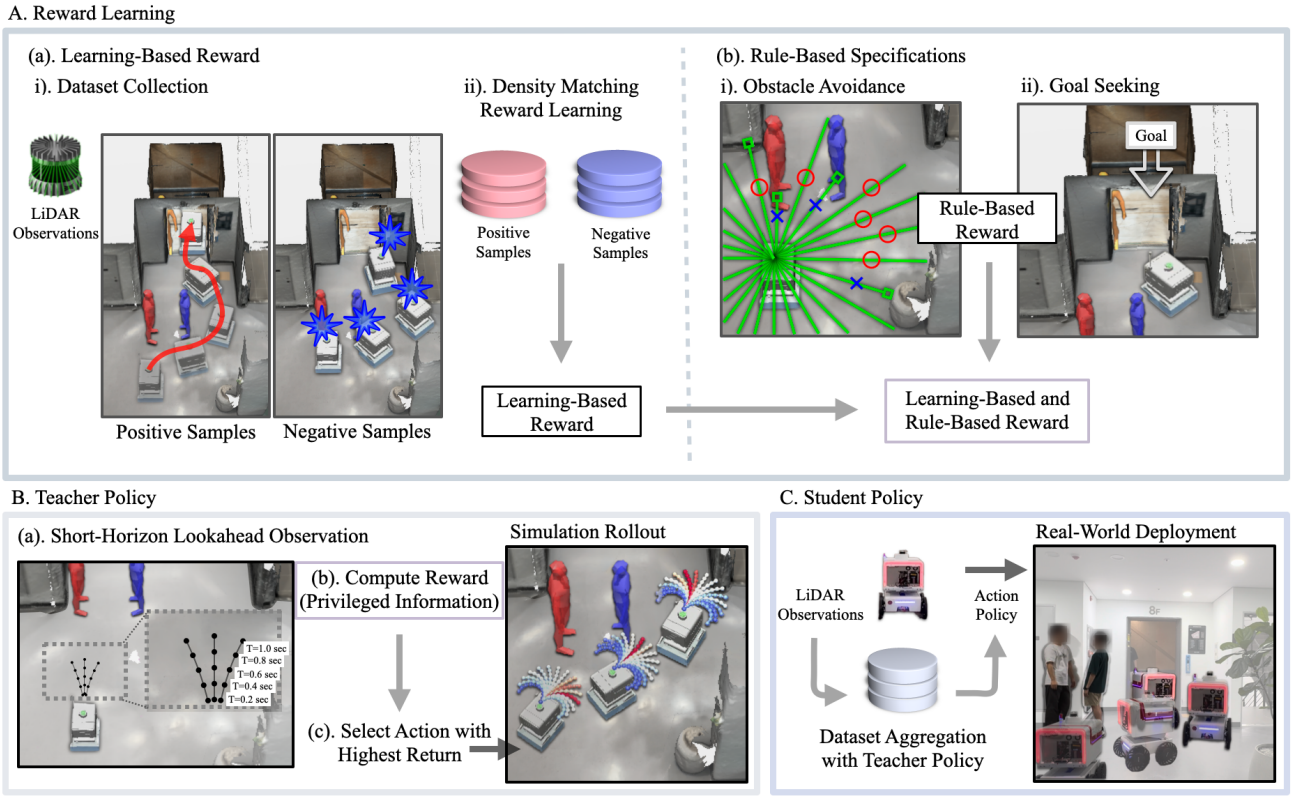


Fig. 1: Overview of the proposed framework. A. Reward learning: (a) density-based reward maps are constructed from positive and negative demonstrations, and (b) augmented with rule-based specifications for obstacle avoidance and goal reaching. B. Teacher policy: short-horizon candidate rollouts are simulated, scored with the combined reward, and used to select safe and adaptive supervisory actions. C. Student policy: the teacher's guidance is distilled into a compact policy conditioned on LiDAR observations, enabling real-world deployment.

Let $K_{UU} \in \mathbb{R}^{N_U \times N_U}$ and $K_{UD} \in \mathbb{R}^{N_U \times N_D}$ have entries $[K_{UU}]_{ij} = k(u_i, u_j)$ and $[K_{UD}]_{ij} = k(u_i, x_j)$. To control reward smoothness and improve numerical stability, two quadratic regularization terms are introduced: one weighted by λ on the RKHS norm of the function, and another weighted by β on the magnitude of the coefficient vector. With $\mathbf{1} \in \mathbb{R}^{N_D}$ the all-ones vector and $\lambda, \beta > 0$, we optimize

$$\max_{\alpha \in \mathbb{R}^{N_U}} \quad \frac{1}{N_D}\, \alpha^\top K_{UD} \mathbf{1} \;-\; \frac{\lambda}{2}\, \alpha^\top K_{UU} \alpha \;-\; \frac{\beta}{2}\, \alpha^\top \alpha, \quad (7)$$

which yields the analytic solution

$$\widehat{\alpha} \;=\; \big(\lambda K_{UU} + \beta I_{N_U}\big)^{-1} \Big(\tfrac{1}{N_D}\, K_{UD}\, \mathbf{1}\Big). \quad (8)$$

To accommodate demonstrations of varying quality [20], each sample is augmented with a fidelity score $\gamma \in [-1, 1]$, where $\gamma \approx +1$ denotes desirable behavior and $\gamma \approx -1$ denotes undesirable behavior. In our implementation, positive demonstrations are assigned $\gamma = +1$ and negative demonstrations are assigned $\gamma = -1$. Positive demonstrations represent successful navigation behaviors that reach the goal while maintaining safe interactions with nearby humans, while negative demonstrations provide complementary information by indicating undesirable state–action patterns associated with unsafe or infeasible interactions. Incorporating both types of examples allows the learned reward to distinguish preferred navigation regions from those that should be avoided.

We employ a smooth leveraged kernel k_{SL} that modulates cross-sample similarity according to these scores:

$$k_{\mathrm{SL}}\big((x,\gamma),(x',\gamma')\big) \;=\; \cos\Big(\tfrac{\pi}{2}\,(\gamma - \gamma')\Big)\; k_{\mathrm{PSD}}(x, x'), \quad (9)$$

where k_{PSD} is chosen as the squared exponential (SE) kernel

$$k_{\mathrm{SE}}(x, x') \;=\; g^2 \exp\Big(-\tfrac{1}{2l^2}\|x - x'\|^2\Big), \quad (10)$$

with hyperparameters g^2 and l^2 controlling the output scale and length-scale. Inducing points $U = \{(u_i, \tilde{\gamma}_i)\}_{i=1}^{N_U}$ and demonstrations $\{(x_j, \gamma_j)\}_{j=1}^{N_D}$ define K_{UU} and K_{UD} via k_{SL} for use in (8), producing a reward estimator that is explicitly sensitive to the distinction between positive and negative samples. The resulting formulation yields a density-based reward map that captures demonstrated navigation preferences while allowing integration with complementary rule-based objectives.

2) Sampling-Based Lookahead Control: At each control cycle, candidate actions are evaluated through short-horizon forward simulation. For a horizon $T > 0$ and discretization step Δt, the rollout of $u \in \mathcal{A}$ produces a state trajectory

$$\Xi(u) = \{x_\ell(u)\}_{\ell=0}^{L}, \qquad L = \big\lfloor T/\Delta t \big\rfloor, \quad (11)$$

where $x_{\ell+1}(u) \;=\; \Phi(x_\ell(u), u, \Delta t)$ with Φ denoting the one-step integrator. Each simulated state is mapped to an

observation (Sec. III-A), yielding the observation rollout $\xi(u) = \{o_\ell(u)\}_{\ell=0}^{L}$.

Each rollout $\xi(u)$ is scored by combining a density-based reward learned from demonstrations with rule-based priors for goal progress and obstacle clearance:

$$R_{\text{PioneeR}}(\xi) = \alpha_{\text{den.}} R_{\text{den.}}(\xi) + \alpha_{\text{goal}} R_{\text{goal}}(\xi) + \alpha_{\text{obs.}} R_{\text{obs.}}(\xi), \tag{12}$$

with nonnegative weights α_*.

The learned reward $R_{\text{den.}}$ evaluates rollouts using a density-based reward map constructed from positive and negative demonstrations:

$$R_{\text{den.}}(\xi) = \tfrac{1}{L} \sum_{\ell=1}^{L} r_{\text{den.}}(o_\ell), \tag{13}$$

where $r_{\text{den.}}(\cdot)$ assigns rewards that reflect both physical occupancy and demonstrated navigation preferences. By leveraging positive demonstrations to indicate desirable behaviors and negative demonstrations to highlight unsafe or undesirable regions, the resulting reward captures aspects of interaction behavior that are difficult to specify analytically while remaining grounded in observed navigation behavior. The goal reward

$$R_{\text{goal}}(\xi) = 1 - \tanh\!\left(\tfrac{d_{\text{goal}}}{d_{\text{total}}}\right), \tag{14}$$

encourages progress toward the target by reducing the remaining distance, while the obstacle reward

$$R_{\text{obs.}}(\xi) = \tanh\!\left(\tfrac{1}{L} \sum_{\ell=1}^{L} d_\ell\right) \tag{15}$$

promotes clearance by favoring trajectories that maintain safe separation from surrounding obstacles.

To balance these components, the weights are modulated as a function of the remaining goal distance. Let

$$r \triangleq \text{clip}\!\left(\tfrac{d_{\text{goal}}}{d_{\text{total}}},\, 0,\, 1\right), \tag{16}$$

and define

$$\alpha_{\text{den.}} = 1 + \cos\big(\pi(1-r)\big), \quad \alpha_{\text{goal}} = 2(1-r), \quad \alpha_{\text{obs}} = 1. \tag{17}$$

The coefficient design serves as a heuristic schedule reflecting that demonstration priors are more informative when the robot is far from the goal, while goal progress becomes increasingly relevant near the target. Since the weights vary smoothly with r, the behavior is expected to change gradually under moderate variations of the coefficient shapes. The formulation assumes goal-conditioned navigation where d_{goal} is available, although alternative scheduling variables could be adopted for tasks without an explicit goal.

In practice, the controller operates at 10 Hz with $T = 3$ s and $\Delta t = 0.3$ s. An exponential moving average with coefficient $\alpha_{\text{EMA}} = 0.5$ is applied to successive commands to reduce abrupt changes. Candidate velocity commands are sampled over a discretized grid, pairing linear velocities $\{0.1, \ldots, 0.8\}$ m/s with 15 uniformly spaced angular velocities in $[-0.4\pi,\ 0.4\pi]$ rad/s. This sampling-based evaluation enables context-aware action selection that reflects both demonstrated behaviors and rule-based safety considerations.

B. Uncertainty-Aware Distillation

The teacher policy relies on privileged information and short-horizon simulation, which makes it unsuitable for direct deployment under real-time constraints. Therefore, we distill the teacher into a compact student policy $\pi_\phi(u \mid o)$ using dataset aggregation [14]. The student is parameterized as a Mixture Density Network (MDN) [24], which models the conditional distribution over velocity commands as a Gaussian mixture:

$$p(u \mid o) \;=\; \sum_{k=1}^{K} \pi_k(o)\, \mathcal{N}\big(u \mid \mu_k(o), \Sigma_k(o)\big), \tag{18}$$

where the mixture weights $\pi_k(o)$, means $\mu_k(o)$, and diagonal covariances $\Sigma_k(o)$ are predicted by a neural network. The network is trained to maximize the likelihood of the teacher's actions, enabling the student to reproduce expert guidance from observations.

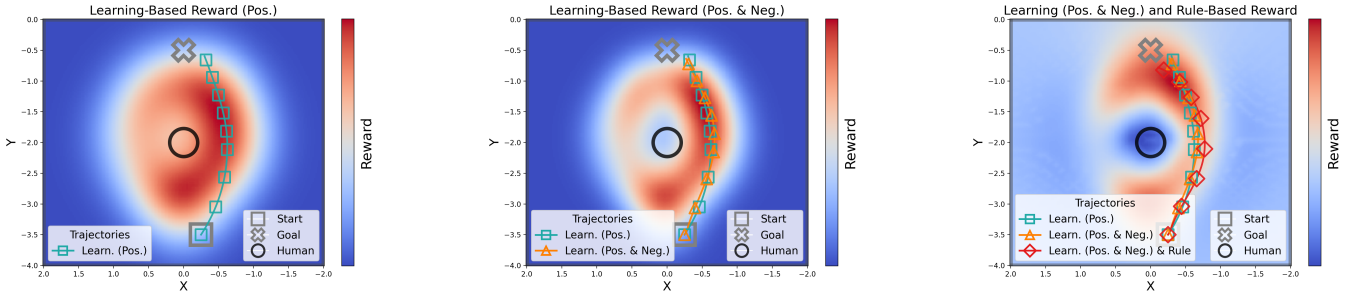
An additional property of the MDN is that it provides closed-form uncertainty estimates via the law of total variance [25]. The predictive covariance decomposes into aleatoric terms, capturing inherent data noise, and epistemic terms, reflecting how each predicted value is different from others:

$$\begin{aligned} \mathbb{V}(y \mid x) &= \mathbb{E}_{k\sim\pi}\big[\mathbb{V}(y \mid x, k)\big] \;+\; \mathbb{V}_{k\sim\pi}\big(\mathbb{E}[y \mid x, k]\big), \\ &= \underbrace{\sum_{j=1}^{K} \pi_j(x)\, \Sigma_j(x)}_{\Sigma_{\text{alea.}}(x)} + \underbrace{\sum_{j=1}^{K} \pi_j(x)\, \big\|\mu_j(x) - \mathbb{E}[y \mid x]\big\|_2^2}_{\Sigma_{\text{epis.}}(x)}. \end{aligned} \tag{19}$$

These uncertainty measures provide informative signals for risk-aware analysis [22], [25], as elevated values may be associated with closer human–robot interactions. In particular, epistemic uncertainty reflects limited model confidence in parts of the observation space that are less represented in the training data. Such situations may arise during unfamiliar or densely interactive encounters, where navigation outcomes are less predictable, making epistemic uncertainty a potential indicator of navigation risk.

V. Experiments

In this section, we present a series of experiments to validate the proposed framework. We first describe the data collection protocol in Sec. V-A, including how positive and negative demonstrations were obtained via keyboard teleoperation. Then, in Sec. V-B, a controlled synthetic study examines the contribution of each component by comparing models trained with positive demonstrations only, with additional negative demonstrations, and with rule-based specifications. Subsequently, Sec. V-C evaluates PioneeR in a realistic elevator co-boarding simulation with humans modeled by a social force model, together with quantitative comparisons against baseline methods. We further analyze uncertainty-aware distillation in Sec. V-C.2, and finally present real-world experiments in Sec. V-D.



(a) (b) (c)

Fig. 2: Reward maps and resulting trajectories with synthetic dataset. (a) Learning-based reward map trained with positive samples, highlighting both feasible corridors. (b) Learning-based reward map trained with positive and negative samples, reducing reward near humans. (c) Reward map combining learning-based and rule-based components, yielding trajectories with greater clearance.

A. Data Collection Protocol

Demonstrations were collected via keyboard teleoperation by an expert operator. Positive demonstrations correspond to successful navigation that reaches the goal while maintaining safe interactions with nearby humans, whereas negative demonstrations were obtained by intentionally recording collisions with humans, providing complementary information about undesirable behaviors. To reduce redundancy, samples were appended only when the robot state changed by more than 5 cm in position or 5° in heading relative to the previous sample, reducing temporal correlation while preserving representative trajectory coverage.

B. Synthetic Study in a Static Environment

We conduct a synthetic study in a planar static setting to illustrate how positive and negative demonstrations influence the learned reward and navigation behavior. For visualization, the reward map is defined over (x, y) coordinates, while trajectories are generated in the full state–action space using the sampling-based lookahead controller. Demonstrations were collected via keyboard teleoperation, yielding 446 positive and 337 negative samples distributed across left and right corridors around a central human.

Figure 2 compares reward maps and trajectories under different conditions. Using only positive demonstrations highlights feasible corridors with a bias toward more frequently demonstrated regions. Incorporating negative demonstrations corresponding to collisions reduces rewards near unsafe regions and shifts the preferred trajectory. Combining the learned reward with rule-based specifications for obstacle avoidance and goal seeking produces trajectories that maintain clearance while achieving the navigation objective.

C. Dynamic Environments in Elevator Co-Boarding

We evaluate PioneeR in a dynamic setting that captures human motion during elevator co-boarding. Humans are simulated using a social force model [26], and the robot receives LiDAR observations $o \in \Omega$ while issuing velocity commands $u = [v, \omega]^T \in \mathcal{U}$. Within PioneeR, a density-based reward learned from demonstrations is combined with rule-based objectives for obstacle avoidance and goal reaching, and the resulting formulation is used by the sampling-based lookahead controller to generate actions. To reflect representative interaction patterns, we consider two human–robot placements: HR-RL (Human-Right, Robot-Left) and HL-RR (Human-Left, Robot-Right).

Training data were collected via keyboard teleoperation for both scenarios, with approximately 732 positive and 170 negative demonstrations in the HR-RL setting, and 662 positive and 194 negative demonstrations in the HL-RR setting. Performance is reported over five random seeds with one hundred trials per seed using success rate (SR), total time (TT), and path length (PL). A trial is considered successful if the robot reaches within 30 cm of the goal within a time limit of 30 s, and TT and PL are computed over successful trials only.

The simulation environment emulates a realistic elevator setting within a $4 \times 4\,\mathrm{m}^2$ map, with geometry modeled after a Schindler 6000 unit [27] (door width $1.8\,\mathrm{m}$, cabin width $2.5\,\mathrm{m}$, depth $2.7\,\mathrm{m}$). Two humans board while one exits, and the robot is initialized between the boarding individuals, creating encounters near the doorway as trajectories cross. In the HR-RL configuration, humans board on the right and the robot goal is on the left, while HL-RR mirrors this arrangement. This scenario induces constrained interactions, crossing pedestrian flows, and limited maneuvering space, providing a structured yet realistic testbed for evaluating socially aware navigation behaviors.

Human motion during elevator co-boarding is modeled using the social force model [26], where each agent moves toward its goal with a preferred velocity while being influenced by repulsive interactions from surrounding agents, walls, and the robot. We assume uncooperative behavior in which humans primarily avoid collisions. Following prior implementations [8], [12], the preferred speed v_{pref} is randomly sampled from $\{0.5, 0.6, 0.7\}$ m/s, and each human is modeled with a radius of $0.6\,\mathrm{m}$.

For comparison, we evaluate two representative baselines. CVaR-BF [8] is an optimization-based controller that combines Conditional Value-at-Risk (CVaR) with control barrier functions to enforce safety through adaptive risk-aware constraints. CrowdNav++ [12] is a learning-based approach that

TABLE I: Performance comparison in two representative human–robot elevator co-boarding scenarios: HR-RL (Human-Right, Robot-Left) and HL-RR (Human-Left, Robot-Right).

Scenario	Method	SR ↑ (%)	TT ↓ (s)	PL ↓ (m)
HR-RL	Ours	99.4	12.24	3.74
	CVaR-BF [8]	72.8	12.82	4.65
	CrowdNav++ [12]	78.6	17.52	4.93
HL-RR	Ours	99.6	12.94	3.88
	CVaR-BF [8]	71.4	12.82	4.75
	CrowdNav++ [12]	64.8	9.71	4.63

models human–robot interactions using a spatio-temporal graph with attention and predicts human intentions within a reinforcement learning framework.

1) Navigation Performance: Table I reports navigation performance in the HR-RL and HL-RR elevator co-boarding scenarios. Across both settings, the proposed method achieves success rates exceeding 99% while maintaining low total time (TT) and path length (PL), indicating reliable and efficient navigation in dynamic human environments. Compared to CVaR-BF [8] and CrowdNav++ [12], our approach achieves substantially higher success while preserving efficiency, highlighting the benefit of integrating demonstration-driven rewards with rule-based safety specifications.

Table II presents ablation results isolating the contributions of individual components. Removing the density-based reward causes the largest drop in success rate, underscoring the importance of demonstration-aligned rewards. Excluding the obstacle prior reduces total time but decreases success, while excluding the goal prior increases total time due to less directed motion. Notably, the density-based reward alone achieves high success rates, while the addition of rule-based priors further improves success rate and efficiency. These trends indicate that learned rewards and rule-based priors play complementary roles, and their combination helps balance adaptability with explicit safety margins.

2) Uncertainty-Aware Distillation: We distill the teacher policy into MDN using the DAgger protocol [14]. Data collection begins with 50 expert episodes, followed by 10 rounds of 50 episodes where the student executes the policy with teacher corrective labels. Training uses a batch size of 256 and a learning rate of 10^{-3}. The MDN consists of two hidden layers with 128 units (ReLU activation, dropout 0.1) and a Gaussian mixture output with $K = 10$ components predicting mixture weights, means, and variances, yielding 26,802 trainable parameters. We also train a standard MLP with the same configuration. As summarized in Table III, the MDN achieves higher success rates than the MLP in both HR-RL and HL-RR scenarios, yielding shorter travel times and path lengths. Representative trajectories of the distilled policy are shown in Fig. 3, demonstrating navigation performance in dynamic elevator co-boarding environments.

Beyond overall performance, we further assess whether the distilled policy can provide signals of risk through predictive uncertainty. Evaluation data are partitioned into

TABLE II: Ablation study of reward terms in two representative human–robot elevator co-boarding scenarios. Each row indicates which reward terms are active (✓) or removed (×).

Scenario	Reward terms			SR ↑ (%)	TT ↓ (s)	PL ↓ (m)
	den.	goal	obs.			
HR–RL	×	✓	✓	83.0	13.61	4.55
	✓	×	✓	99.4	15.70	3.79
	✓	✓	×	94.8	11.64	3.72
	✓	×	×	96.6	14.65	3.78
HL–RR	×	✓	✓	79.8	13.37	4.54
	✓	×	✓	99.4	16.48	3.92
	✓	✓	×	98.8	12.33	3.86
	✓	×	×	99.4	15.71	3.90

safe and risky frames according to the minimum human-robot distance: frames with clearance greater than 0.8 m are labeled safe, while those below this threshold are labeled risky. In HR-RL, this results in 62,374 safe frames and 5,467 risky frames, and in HL-RR, 67,806 safe frames and 3,245 risky frames. Across both scenarios, epistemic and aleatoric uncertainty values are consistently higher in risky frames, indicating that the MDN provides informative signals about potential hazards. Such estimates can guide conservative fallback strategies such as reverting to rule-based control [25], thereby complementing the efficiency of the distilled policy with explicit safety awareness.

D. Real-World Demonstrations

To demonstrate the practicality of the proposed framework, we conducted real-world experiments in two representative elevator co-boarding scenarios with human participants, as illustrated in Fig. 4. A four-wheel mobile robot executed control at 10 Hz, with policy observations derived from 3D LiDAR data converted into 2D laser scans. In the first scenario, a single policy was deployed across varied co-boarding situations, adapting to differences in human motion. In the second scenario, the robot navigated encounters involving multiple humans and managed the interactions effectively. These trials demonstrate that the proposed framework extends beyond simulation and supports reliable operation in real-world human–robot interaction settings.

E. Limitations

While the proposed framework demonstrates strong performance in both simulation and real-world experiments, certain limitations remain. First, the simulation of human motion is based on a social-force model, which provides structured scenarios but does not fully capture the variability of real pedestrian behaviors. Second, although the reward learning formulation supports a continuous fidelity variable $\gamma \in [-1, 1]$ to represent a spectrum of demonstration quality, in practice we restricted γ to binary labels for positive and negative examples. This simplification improves training efficiency but may limit the ability to capture more nuanced variations in demonstration quality. Third, the density-based reward relies on kernel hyperparameters, and while empirically stable behavior was observed, a more systematic anal-

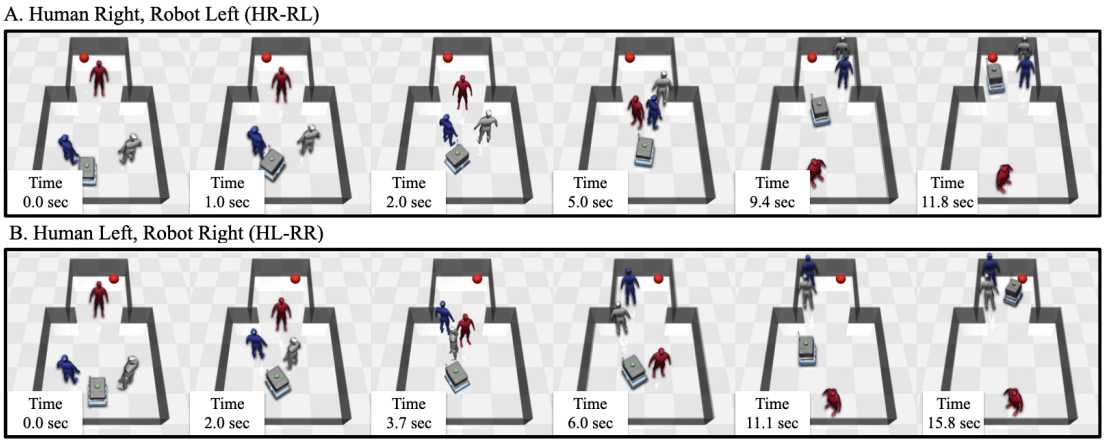


Fig. 3: Simulation snapshots of elevator co-boarding scenarios. (A) HR-RL: Human on the right, Robot on the left. (B) HL-RR: Human on the left, Robot on the right.

TABLE III: Comparison of distillation performance from the teacher policy across different network architectures.

Scenario	Method	SR ↑ (%)	TT ↓ (s)	PL ↓ (m)
HR-RL	PioneeR-MDN	98.0	13.67	4.62
	PioneeR-MLP	95.8	15.02	6.49
HL-RR	PioneeR-MDN	100.0	14.16	4.70
	PioneeR-MLP	98.8	15.29	4.78

TABLE IV: Mean uncertainty values for safe–risky frames.

Scenario	Safe		Risky	
	Epi.	Ale.	Epi.	Ale.
HR-RL	0.290	0.310	0.382	0.407
HL-RR	0.342	0.270	0.523	0.426

ysis of parameter sensitivity could provide additional insight into robustness. Fourth, the real-world experiments primarily demonstrate feasibility, and more extensive quantitative evaluation with larger datasets and repeated trials would further strengthen empirical validation. Finally, although the proposed framework allows additional social objectives through rule-based reward components, our evaluation focuses on navigation success, efficiency, and safety in constrained interactions. Incorporating broader social compliance metrics could provide further insight into behavior quality.

VI. Conclusion

This paper presented a navigation framework that integrates density-based reward learning from positive and negative demonstrations with rule-based objectives for obstacle avoidance and goal reaching. The teacher policy generates supervision under this reward formulation, and a student policy is distilled for real-time deployment with predictive uncertainty. Experiments in synthetic and dynamic elevator co-boarding scenarios demonstrated improved performance compared to baselines, and real-world trials confirmed feasibility on a mobile robot operating alongside humans.

Acknowledgment

This work was supported by Mobinn (20%). This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), 20%); (No. RS-2022-II220612, Geometric and Physical Commonsense Reasoning based Behavior Intelligence for Embodied AI, 20%); and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00211357, Smart Assembler: Robot Active Learning for Unseen Parts Assembly, 20%). This research was also financially supported by the Ministry of Trade, Industry, and Energy (MOTIE), Korea, under the "Global Industrial Technology Cooperation Center program" supervised by the Korea Institute for Advancement of Technology (KIAT) (Grant No. P0028435, 20%).

References

[1] P. T. Singamaneni, P. Bachiller-Burgos, L. J. Manso, A. Garrell, A. Sanfeliu, A. Spalanzani, and R. Alami, "A survey on socially aware robot navigation: Taxonomy and future challenges," *The International Journal of Robotics Research*, vol. 43, no. 10, pp. 1533–1572, 2024.

[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*, vol. 12, no. 3, pp. 1–39, 2023.

[3] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE robotics & automation magazine*, vol. 4, no. 1, pp. 23–33, 2002.

[4] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," *The international journal of robotics research*, vol. 5, no. 1, pp. 90–98, 1986.

[5] J. Van Den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics Research: The 14th International Symposium ISRR*. Springer, 2011, pp. 3–19.

[6] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *2019 18th European control conference (ECC)*. Ieee, 2019, pp. 3420–3431.

[7] W. Xiao, N. Mehdipour, A. Collin, A. Y. Bin-Nun, E. Frazzoli, R. D. Tebbens, and C. Belta, "Rule-based optimal control for autonomous driving," in *Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems*, 2021, pp. 143–154.

Fig. 4: Real-world elevator co-boarding experiments with human participants. (A) Single policy across diverse scenarios. (B) Scenario involving multiple human interactions.

[8] X. Wang, T. Kim, B. Hoxha, G. Fainekos, and D. Panagou, "Safe navigation in uncertain crowded environments using risk adaptive cvar barrier functions," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025.

[9] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1343–1350.

[10] L. Liu, D. Dugas, G. Cesari, R. Siegwart, and R. Dubé, "Robot navigation in crowded environments using deep reinforcement learning," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5671–5677.

[11] Z. Xie, P. Xin, and P. Dames, "Towards safe navigation through crowded dynamic environments," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4934–4940.

[12] 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)*. IEEE, 2023, pp. 12 015–12 021.

[13] M. Pfeiffer, S. Shukla, M. Turchetta, C. Cadena, A. Krause, R. Siegwart, and J. Nieto, "Reinforced imitation: Sample efficient deep reinforcement learning for mapless navigation by leveraging prior demonstrations," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4423–4430, 2018.

[14] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.

[15] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 1.

[16] K. Bektaş and H. I. Bozma, "Apf-rl: Safe mapless navigation in unknown environments," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7299–7305.

[17] S. Dey, A. Sadek, G. Monaci, B. Chidlovskii, and C. Wolf, "Learning whom to trust in navigation: dynamically switching between classical and neural planning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 5235–5242.

[18] Y. Zhu, Z. Wang, C. Chen, and D. Dong, "Rule-based reinforcement learning for efficient robot navigation with space reduction," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 2, pp. 846–857, 2021.

[19] K. Long, C. Qian, J. Cortés, and N. Atanasov, "Learning barrier functions with memory for robust safe navigation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4931–4938, 2021.

[20] S. Choi, K. Lee, and S. Oh, "Robust learning from demonstration using leveraged gaussian processes and sparse-constrained optimization," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1152–1159.

[21] ——, "Robust learning from demonstrations with mixed qualities using leveraged gaussian processes," *IEEE Transactions on Robotics*, vol. 35, no. 3, pp. 564–576, 2019.

[22] J. Oh, G. Lee, J. Park, W. Oh, J. Heo, H. Chung, D. H. Kim, B. Park, C.-G. Lee, S. Choi *et al.*, "Towards defensive autonomous driving: Collecting and probing driving demonstrations of mixed qualities," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 12 528–12 533.

[23] S. Choi, K. Lee, H. A. Park, and S. Oh, "Density matching reward learning," *arXiv preprint arXiv:1608.03694*, 2016.

[24] C. M. Bishop, "Mixture density networks," 1994.

[25] S. Choi, K. Lee, S. Lim, and S. Oh, "Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6915–6922.

[26] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[27] S. Group, "Schindler 6000," https://www.jardineschindler.com/en/elevators/passenger/schindler-6000.html, 2025, accessed: 2025-08-31.