Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Robotics

Perturbation-Resilient Autonomous Navigation with Distributionally Robust Reinforcement Learning

Zhaofan Zhang[1] , **Minghao Yang**[1] , **Sihong Xie**[1*] and **Hui Xiong**[1,2*]
[1]Thrust of Artificial Intelligence, The Hong Kong University of Science and Technology (Guangzhou), China
[2]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology Hong Kong SAR, China
{zzhang398, myang272}@connect.hkust-gz.edu.cn, sihongxie@hkust-gz.edu.cn, xionghui@ust.hk

Abstract

The robustness of autonomous vehicles such as drones and Unmanned Surface Vehicles (USV) is crucial when facing unknown and complex marine environments, especially when heteroscedastic observational noise poses significant challenges to sensor-based navigation tasks. Recently, Distributional Reinforcement Learning (DistRL) has shown promising results in some challenging autonomous navigation tasks without prior environmental information. However, these methods overlook situations where noise patterns vary across different environmental conditions, hindering safe navigation and disrupting the learning of value functions. To address the problem, we propose DRIQN to integrate Distributionally Robust Optimization (DRO) with implicit quantile networks to optimize worst-case performance under natural environmental conditions. Leveraging explicit subgroup modeling in the replay buffer, DRIQN incorporates heterogeneous noise sources and target robustness-critical scenarios. Experimental results based on the risk-sensitive environment demonstrate that DRIQN significantly outperforms state-of-the-art methods, achieving +13.51% success rate, -12.28% collision rate and +35.46% for time saving, +27.99% for energy saving, compared with the runner-up.

1 Introduction

Autonomous navigation is a foundational capability for unmanned systems, enabling both USVs and drones to transition from teleoperation to autonomous, goal-directed operation under uncertain and rapidly changing conditions. Reliable navigation empowers vehicles to undertake long-duration missions ranging from port oil spill cleanup [Elmakis and Degani, 2023] and environmental monitoring [Maharjan *et al.*, 2022] to border security [Fernandes *et al.*, 2025] and search-and-rescue [Wang *et al.*, 2023b] while simultaneously reducing operational risk, human exposure, and overall cost. In these navigation tasks, the environment is characterized by complexity and partial observability. The intricate variations in water currents and airflow are difficult to monitor in real-time, and obstacles may also be encountered [Lolla *et al.*, 2015]. Moreover, the perception devices used by the robot may face various unobservable interferences within the environment, resulting in noisy observations [Vieira *et al.*, 2020]. Specifically, heterogeneous extreme natural conditions (e.g., weather, sea state, geographic locality) induce trajectory-specific, non-stationary observational noise. As illustrated in Figure 1, the observational perturbation exacerbates the inherent uncertainty of complex marine environments and impedes accurate risk-sensitive policy learning, creating substantial navigation hazards for unmanned vehicles, whereas our proposed DRIQN remains robust to such corruption and progresses reliably toward the target.

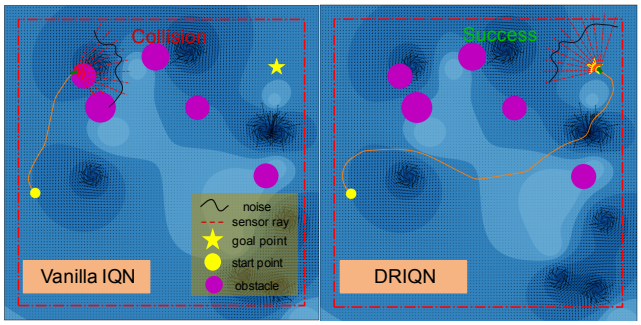

Figure 1: **IQN vs. Robust DRIQN:** USV navigation trajectories visualization with observational noise. Vanilla IQN may be biased by noisy observations, resulting in risky policies.

Prior works on autonomous navigation can be mainly grouped into: (i) *model-based* methods [Wang *et al.*, 2023a; Asfihani *et al.*, 2025]; (ii) models with *prior-dependent* knowledge [Han *et al.*, 2016; Hong *et al.*, 2024]; (iii) *direct observation-driven decision-making* approaches (e.g., APF [Sun *et al.*, 2019], BA [McGuire *et al.*, 2018] and reinforcement learning [Lin *et al.*, 2023; Wang *et al.*, 2024]). Direct observation-driven approaches make decisions based on sensor observation without prior environmental knowledge and extra modeling design. Despite promising results, model-based approaches (e.g., nonlinear MPC) incorporate analytical vehicle dynamics into optimal control, requiring high-fidelity models and suffering degradation under uncer-

*Corresponding authors.

tainties. Prior-dependent methods further face limitations in prior availability and freshness, incurring substantial precomputation overhead. In contrast, observation-driven decision-making methods offer a generalizable, prior-independent solution for navigation in unknown environments, albeit with heightened sensitivity to observational noise. Recently, as one kind of prominent observation-driven decision-making method, Deep Reinforcement Learning (DRL) has emerged for robotic navigation in complex environments [Hossain *et al.*, 2024]. Building on this advantage, some DistRL methods [Lin *et al.*, 2023; Liu *et al.*, 2023] enable risk-sensitive navigation through quantile-based return distribution modeling. By employing task-specific reward designs and distortion functions with Conditional Value-at-Risk (CVaR), DistRL reshapes return distributions to focus on critical lower-tail outcomes while constraining distributional support, thereby enhancing robust policy learning. Nevertheless, without environmental priors or modeling knowledge for constraint factors, these approaches cannot be readily optimized via constructing robust mechanisms, thereby lacking explicit safety guarantees under heterogeneous noise-corrupted observations. This presents a pressing challenge: how can we ensure the safety of unmanned surface navigation in multi-noise-corrupted marine environment just using a data-driven optimization approach?

To address the challenge, we propose Distributionally Robust Implicit Quantile Networks (DRIQN), a novel framework ensuring risk-sensitive policies via quantile regression under observational noise. Leveraging distributionally robust optimization, DRIQN reformulates the DRO problem as a semi-definite program (Eq. 13) to holistically address diverse noise influences. Crucially, DRIQN implements "gradient substitution" replacing the network's output-layer gradients with DRO-based gradients to enhance robustness without auxiliary modules. Considering the scalability challenge of DRO caused by the large-scale datapoint-level optimization process, we transform standard datapoint-based DRO by leveraging transition data's inherent structure: under specific environmental conditions, noisy transitions naturally form homogeneous subgroups following corresponding noise patterns. This structural property enables reformulating the original datapoint-level DRO problem as the supremum over subgroup-level loss functions. Each subgroup represents unlabeled transition data clusters with distinct noise characteristics. Experimental results demonstrate that DRIQN significantly outperforms baselines: achieving +13.51% success rate, -12.28% collision rate and +35.46% for time saving, +27.99% for energy saving, compared with the runner-up. To our knowledge, DRIQN is the first method tackling multiple observational noise patterns from natural conditions, enabling robust risk-sensitive policies for autonomous navigation.

The main contributions of this paper are as follows:

- We propose a novel, streamlined distributionally robust optimization framework that enhances RL agent performance against perturbed observations through substitution of DRO-based gradients.
- We present a risk-sensitive and robust navigation approach for developing an online path planner by unifying DistRL and DRO, maintaining robustness and safety in the presence of complex environmental noise patterns.
- Simulated experiments show that the resulting planner remains robust to observation perturbations, enabling safe goal reaching under strong disturbances, unknown flows, and obstacles, and that this robustness generalizes across tasks and vehicle types.

2 Related Work

Among autonomous navigation paradigms (model-based, prior-dependent, and direct observation-driven), observation-driven methods are particularly suitable for planning in unknown environments due to their generality and minimal prior requirements. Classic representatives include Artificial Potential Field (APF) methods [Khatib, 1985; Sun *et al.*, 2019], which combine goal attraction and obstacle repulsion but suffer from local minima and goal-proximity failures, and Bug algorithms (e.g., Bug1/Bug2) [Lumelsky and Stepanov, 1986; McGuire *et al.*, 2018], which follow obstacle boundaries for reactive planning yet often yield highly suboptimal paths due to exhaustive perimeter traversal. More recently, deep RL has been widely adopted for navigation. For safety-critical settings, safe RL methods [Lütjens *et al.*, 2018; Dawood *et al.*, 2024; Tomilin *et al.*, 2025] enforce constrained optimization, but typically require handcrafted constraints and can still incur catastrophic failures under transient violations. However, adversarial robustness to synthetic perturbations and constraint-based Safe RL are not well aligned with our setting, which focuses on naturally occurring observation noise without manually imposed safety constraints. Without the aforementioned limitations, DistRL demonstrated superior performance in risk-sensitive tasks such as quadruped locomotion [Shi *et al.*, 2024a], pursuit-evasion [Zhang *et al.*, 2025], and autonomous navigation [Lin *et al.*, 2023; Liu *et al.*, 2023]. Nevertheless, severe observational noise induces a distributional shift between the learned and deployed returns. This mismatch compromises the reliability of standard distribution estimates.

For this non-trivial challenge, DRO can provide a principled framework for robust decision-making under uncertainty [Nietert *et al.*, 2023]. Specifically, DRO solves a minimax problem that minimizes the maximum expected loss over an uncertainty set of plausible data distributions centered on the empirical distribution [Staib and Jegelka, 2019]. Unlike expected-value optimization under domain randomization [Tobin *et al.*, 2017], which does not directly control worst-case degradation, DRO emphasizes adverse extremes, aligning more closely with navigation safety. Prior studies have leveraged DRO to improve robustness, but many introduce auxiliary components (e.g., generative models [Shi *et al.*, 2024b; Ramesh *et al.*, 2024]), model-based rule [Shi and Chi, 2024; He *et al.*, 2025], offline data [Leung *et al.*, 2025]). These requirements limit applicability to real-world problem, where agents must learn online in unknown environments without prior models or curated datasets.

3 Methodology

3.1 Problem Overview

An autonomous vehicle does not have full knowledge of the environment, but instead receives noisy partial observations of robot's surroundings constrained by the limited sensing range of its sensors. We formulate the automated navigation problem as a *perturbed* Partially Observable Markov Decision Process $\widetilde{\mathcal{M}}$, denoted as $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, R, \gamma, \nu)$, where $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{O}$ represent the state, action and observation space, respectively. At each time step t, the agent receives a corrupted observation $\nu(o_t)$ subject to natural disturbances, where $o_t \in \mathcal{O}$ denotes the observation of current state s_t. Based on this observation, the agent selects an action $a_t \in \mathcal{A}$ according to policy $\tilde{\pi}(a_t \mid \nu(o_t))$. This leads the agent to transition to the next state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$, and the agent receives a reward $r_t \sim R(s_t, a_t)$, as well as a perturbed observation $\nu(o_{t+1})$ of s_{t+1}. For observation-disturbed policy $\tilde{\pi}$, the discounted sum of future rewards is a random variable $Z^{\tilde{\pi}}(s_t, a_t)$, defined as $Z^{\tilde{\pi}}(s_t, a_t) = \sum_{k=0}^{\infty} \gamma^k R(s_{t+k}, a_{t+k})$, where $\gamma \in (0, 1)$ denotes the future reward discount factor. The objective of standard RL under noisy observations is to maximize the expected discounted return, formalized by the value function: $Q^{\tilde{\pi}}(s_t, a_t) = \mathbb{E}\left[Z^{\tilde{\pi}}(s_t, a_t)\right]$. Accordingly, expectations are risk-neutral, representing returns solely through $Q^{\tilde{\pi}}(s, a)$. DistRL addresses the limited expressiveness of this expectation by estimating the distribution of $Z^{\tilde{\pi}}(s, a)$. Environmental noise is incorporated into the state transition kernel to form the state representation.

3.2 Distributional Reinforcement Learning

Distributional Reinforcement Learning explicitly models the return distribution Z governed by the following *distributional* Bellman equation and optimality equation:

$$Z^{\pi}(s, a) \overset{D}{=} R(s, a) + \gamma Z^{\pi}(s', a'), \tag{1}$$

$$\mathcal{T}Z^{\pi}(s, a) \overset{D}{=} R(s, a) + \gamma Z^{\pi}\left(s', \underset{a'}{\operatorname{argmax}}\, \mathbb{E}[Z^{\pi}(s', a')]\right), \tag{2}$$

where $\mathcal{T}$ represents the *distributional Bellman optimality operator*. Here $\overset{D}{=}$ indicates distribution equality.

We utilize DistRL with IQN as a risk-sensitive path planner that has achieved promising performance in many tasks of robotics [Liu *et al.*, 2023; Lin *et al.*, 2023; Shi *et al.*, 2024a]. IQN utilizes the quantiles to provide a flexible representation of return distribution Z, denoted as $Z_{\tau} := F_Z^{-1}(\tau)$, where $\tau \sim U([0, 1])$ and F^{-1} is inverse cumulative distribution function. To effectively capture risk-sensitive behaviors, IQN incorporates a distortion function β to adjust the quantile distribution, emphasizing certain regions of the return distribution. Then the distorted expectation of Z under $\beta : [0, 1] \to [0, 1]$ is defined by:

$$Q_{\beta}(s, a) = \mathbb{E}_{\tau \sim U([0,1])}[Z_{\beta(\tau)}(s, a)]. \tag{3}$$

The standard deep Q-Learning framework trains a neural network-based state-action value function by iteratively minimizing a loss function, which is defined through the temporal difference (TD) error. The TD error is defined as:

$$\delta = r + \gamma \max_{a'} Q(s', a') - Q(s, a). \tag{4}$$

The IQN loss function integrates the sampled temporal difference error Eq. (6) with the *Huber* quantile regression [Huber, 1992] Eq. (7) to optimize the quantile value network:

$$\mathcal{L}_{IQN}(s, a, r, s') = \frac{1}{N'} \sum_{i=1}^{N} \sum_{j=1}^{N'} \rho_{\tau_i}^{\kappa}(\delta^{\tau_i, \tau_j'}) \tag{5}$$

where N and N' denote the counts of i.i.d. samples $\{\tau_i\}_{i=1}^{N}$ and $\{\tau_j'\}_{j=1}^{N'}$ from $\mathcal{U}([0, 1])$. The temporal difference error between quantile targets is defined as:

$$\delta^{\tau_i, \tau_j'} = r + \gamma Z_{\tau_j'}(s', a') - Z_{\tau_i}(s, a) \tag{6}$$

To mitigate the impact of outliers, we employ the Huber quantile loss:

$$\begin{aligned} \rho_{\tau}^{\kappa}(u) &= |\tau - \mathbb{I}_{\{u<0\}}| \left(\frac{\mathcal{L}_{\kappa}(u)}{\kappa}\right) \\ \mathcal{L}_{\kappa}(u) &= \begin{cases} \frac{1}{2}u^2, & |u| \le \kappa \\ \kappa(|u| - \kappa/2), & \text{otherwise} \end{cases} \end{aligned} \tag{7}$$

Minimizing the expectation of this loss yields a convex objective for the quantile-based method. Consequently, the greedy policy is approximated by averaging over K quantile samples $\tilde{\tau}_k \sim \mathcal{U}([0, 1])$:

$$\tilde{\pi}_{\beta}(s) = \underset{a}{\operatorname{argmax}} \frac{1}{K} \sum_{k=1}^{K} Z_{\beta(\tilde{\tau}_k)}(s, a) \tag{8}$$

with actions selected through $\tilde{\pi}_{\beta}(s)$.

3.3 Navigation under Observational Perturbation

To evaluate navigation robustness, we utilize two distinct simulation environments: a drone navigating in cluttered indoor spaces and a USV navigating dynamic ocean scenarios. In both settings, the agent is tasked with reaching a target destination under natural observational noise and environmental disturbances. The RL setting is defined for the marine environment and is applied analogously to the drone environment with only minor adjustments following prior work [Liu *et al.*, 2022]: **Reward.** To encourage goal-directed motion and collision avoidance, we define $r_t = r_{\text{step}} + \alpha(d_{t-1} - d_t) + r_{\text{collision}}\mathbb{I}_{\text{collision}} + r_{\text{goal}}\mathbb{I}_{\text{goal}}$, where d_t is the distance to the goal at time t, and $\mathbb{I}_{\text{collision}}$ (resp. $\mathbb{I}_{\text{goal}}$) equals 1 if a collision occurs (resp. the goal is reached) at t, and 0 otherwise. We set $r_{\text{step}} = -1.0$, $r_{\text{collision}} = -50.0$, $r_{\text{goal}} = 100.0$, and $\alpha = 1.0$ following [Liu *et al.*, 2022]. **Observation.** Following real-world sensing, the observation at time t is $O_t = (O_{\text{velocity}}, O_{\text{goal}}, O_{\text{LiDAR}})$, where O_{velocity} is the seafloor-relative velocity, O_{goal} is the goal position in the robot frame, and O_{LiDAR} denotes the LiDAR return vector that captures obstacle reflections within $10\,\text{m}$ over a $2\pi/3\,\text{rad}$ field of view. **Action.** The action at time t is $a_t = (a(t), w(t))$, where $a(t)$ controls the change rate of speed magnitude and $w(t)$ controls the change rate of heading direction. The discrete action set is $a \in \{-0.4, 0.0, 0.4\}\, m/s^2$ and $w \in \{-0.52, 0.0, 0.52\}\, rad/s$, with forward speed constrained to $[0, v_{\max}]$. **Observational Perturbation.** To model natural noise induced by meteorological, illumination, and environmental factors (rather than adversarial attacks), we use non-synthetic noise sources specified in Section 4.1.

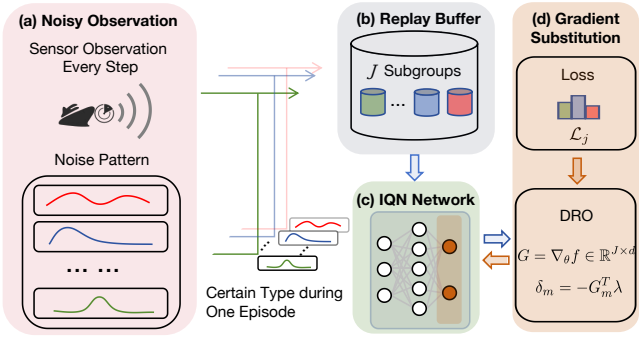

Figure 2: **Overall framework of our method.** In (a), the USV receives noisy observations specific to its current natural condition, assuming a consistent noise pattern within one episode due to proximity of environmental condition. Therefore, the replay buffer in (b) accumulates subgroups with distinct noise patterns. This variation has a significant impact on the training of the implicit quantile networks in (c). To enhance robustness, DRIQN employs "Gradient Substitution" in (d), replacing the gradients of the output layer with those computed via DRO.

3.4 DRO for Online DistRL

IQN has demonstrated superior performance as a risk-sensitive policy for path planning in an ideal environment. However, the planner relies on online distributional reinforcement learning training without prior information, meaning that errors in actions under noisy observations will cause reward signal mismatching and policy deviation. This issue is exacerbated by the presence of multiple noise subgroups within the replay buffer. Despite its impact, this problem remains largely unexplored in existing research.

To address heterogeneous noise in robot sensor observations, we propose DRIQN, as shown in Figure 2. DRIQN casts robustness under noisy transition data as an optimization problem and incorporates DRO during training.

Based on prior work on DRO theory [Qi *et al.*, 2020; Shen *et al.*, 2022], we further integrate the DRO objective into IQN training by formulating the overall optimization target as:

$$\min_{\theta\in\mathbb{R}^d} F_p(\theta) = \min_{\theta\in\mathbb{R}^d} \sup_{\mathbf{P}\in\mathbb{R}^n} \sum_{i=1}^{n} p_i \mathcal{L}_{IQN}(T_i) - d(\mathbf{P}, \mathbf{1}/n), \tag{9}$$

where $\mathcal{L}_{IQN}$ denotes the IQN loss function applied to the noisy transition datapoint $T_i = (s_i,\, a_i,\, r_i,\, s_{i+1})$ from replay buffer of size n. The IQN model uses parameters $\theta \in \mathbb{R}^d$, while $\mathbf{P} = (p_1, \ldots, p_n)$ denotes noise-characterizing density ratios satisfying $\sum p_i = 1$ and $p_i > 0$. $d(\mathbf{P}, \mathbf{1}/n)$ quantifies the difference between $\mathbf{P}$ and the uniform probabilities $\mathbf{1}/n$.

To adapt the DRO to big data and large parameter scenario, we focus on the uncertainty set $\mathcal{Q}$, which is partitioned into J subgroups that represent the subgroups of transition data with different noise types. Based on this replay buffer partition, we propose an efficient gradient-based learning approach to solve the optimization problem.

Given identical noise types per subgroup as detailed in Section 4.1, we implement independent sampling across all replay-buffer subgroups. Following Group DRO [Liu *et al.*, 2025], we optimize for worst-group performance by considering an ambiguity set formed by a finite set of subgroup distributions and minimizing the maximum subgroup loss. While this worst-group objective improves robustness, it can be overly conservative and may under-utilize informative samples from non-worst groups. To further improve both comprehensive robustness and data utilization efficiency beyond the worst-case subgroup, we compute the optimization gradient as a convex combination of gradients from all noisy subgroups. The optimization gradient in DRIQN is used to replace the original gradient during the backpropagation through the last quantiles layer of the IQN network.

We formalize the inner term of DRO problem Eq. (9) with a partitioned uncertainty set $\mathcal{Q}$. For simplicity, we construct $\mathcal{Q}$ through considering limited J kinds of noise, i.e. $\mathcal{Q} := \{q_1, q_2, ..., q_J\}$.

$$\sup_{\mathbf{P}\in\mathbb{R}^J, q\in\mathcal{Q}} \sum_{j=1}^{n} p_j \mathbb{E}_{(T_j)\sim q_j} \left[\mathcal{L}_{IQN}(T_j)\right] - d(\mathbf{P}, \mathbf{1}/J). \tag{10}$$

To ensure a fair comparison of natural-condition-induced noise types, we assume equiprobable occurrence for all categories, which yields $d(\mathbf{P}, \mathbf{1}/J) = 0$ in the final term. So the inner term of DRO will be expressed as:

$$\sup_{q\in\mathcal{Q}} \mathbb{E}_{T\sim q} \left[\mathcal{L}_{IQN_q}(T)\right]. \tag{11}$$

Now we go back to the complete DRO minimax term Eq. (9). To solve the problem, we aim to minimize the expected loss over the uncertainty set $\mathcal{Q}$ following:

$$\min_{\theta} \max_{1\le j\le J} \frac{1}{c_j} \sum_{k=1}^{c_j} \mathcal{L}_{IQN}(T_{jk}), \tag{12}$$

where J is the number of noise subgroups, c_j is the number of samples in subgroup j that has unique noise type, T_{jk} refers to the k-th transition data in the j-th subgroup. Our method provides a more scalable and efficient way to obtain the gradient descent direction by solving optimization problem Eq. (12) through learning the supremum of the set of loss functions with respect to each noise subgroup.

As the optimization problem described in Eq. (12), the right side noted as $f_j(\theta) = \frac{1}{c_j}\sum_{k=1}^{c_j} \mathcal{L}_{IQN}(T_{jk})$ during the iteration m can be linearized to achieve convex approximation, therefore we rewrite the right side as $f_j(\theta_m) + \langle \nabla f_j(\theta_m), \theta - \theta_m \rangle$.

Since the inner term for $\max$ operation has been smoothed by linearization, to ensure finding the stable descent direction during the $\min$ stage when $\max_{1\le j\le J} f_j(\theta_m) + \langle \nabla f_j(\theta_m), \theta - \theta_m \rangle$ is not strictly convex with respect to parameter θ. Thus, we add a regularization term $\|\theta - \theta_m\|_2$. The gradient descent direction is set as $\delta := \theta - \theta_m$, we obtain the equivalent form of the DRO problem:

$$\min_{\delta,\kappa} \|\delta\|_2 + \kappa, \tag{13a}$$

$$\text{s.t. } f_j(\theta_m) + \langle \nabla f_j(\theta_m), \delta \rangle \le \kappa,\ \forall 1 \le j \le J. \tag{13b}$$

Theorem 1 (Equivalent dual quadratic program)**.** *Let $\theta_m \in \mathbb{R}^d$ be the model parameter at iteration m and define the subgroup losses $f_j(\theta)$, $j = 1, \ldots, J$. Denote $G_m := \nabla_\theta f(\theta_m) \in$*

$\mathbb{R}^{J\times d}$, whose j-th row is $\nabla_\theta f_j(\theta_m)^\top$, and consider the quadratic program in Eq. (13). The unique optimal descent direction of this problem is a convex combination form:

$$\delta_m^\star = -G_m^\top \lambda^\star, \quad \lambda^\star \in \Delta_J := \Big\{\lambda \in \mathbb{R}^J_{\geq 0} \;\Big|\; \sum_{j=1}^{J} \lambda_j = 1\Big\}, \tag{14}$$

where $\lambda^\star$ solves the dual quadratic program

$$\min_{\lambda\in\Delta_J} \; \tfrac{1}{2}\,\lambda^\top G_m G_m^\top \lambda - \lambda^\top f(\theta_m). \tag{15}$$

Hence $\delta_m^\star$ is a linear combination of all subgroup gradients at iteration m.

The formulation preserves the Bellman operator's monotonicity and γ-shift invariance, ensuring that value evaluation is well-posed for any stationary policy. Building on this foundation, Theorem 1 (proof adapted from [Shen *et al.*, 2022] is detailed in Section A of Supplementary Material) establishes that the original constrained minimax problem can be solved by an equivalent dual quadratic program of dimension J (the number of noise subgroups). This equivalence allows the optimization to be reduced from searching over high-dimensional vectors to finding a linear combination of subgroup gradients, incurring a computational complexity of $O(J^3)$. Crucially, the resulting convex-combination coefficients serve as a weight vector that adaptively allocates importance across subgroups. This implies that DRIQN leverages a weighted combination of subgroup gradients rather than discarding all but the worst-case sample. With bounded mini-batch gradients of $\max_{1\leq j\leq J} f_j(\theta_m) + \langle \nabla f_j(\theta_m), \theta - \theta_m\rangle$, the stochastic convergence rate is $\mathcal{O}(T^{-1/2})$ [Hazan *et al.*, 2006], where T is the number of iterations.

Remark. *As characterized in Theorem 1, standard IQN updates suffer from gradient conflict under heterogeneous noise, where gradients from benign and hazardous subgroups oppose each other (i.e., $\langle \nabla f_j, \nabla f_k\rangle < 0$). This cancellation drives the learned quantiles toward the average return distribution, causing the agent to significantly underestimate the tail risks associated with worst-case scenarios.*

4 Experiments

4.1 Experimental Settings

Simulation Environment. Following recent work, we evaluate robustness in simulation, as real-world marine experiments are prohibitively costly. To better approximate real-world navigation under uncertainty, we evaluate our method in two perturbation-prone simulation settings. We first adopt a risky marine environment for USV navigation based on prior work [Lin *et al.*, 2023]. To further assess generalization across robot types and navigation contexts, we additionally consider a UAV task in an indoor cluttered environment [Liu *et al.*, 2022]. In both settings, we use a purely sensor-based navigation scheme without any prior knowledge. For the marine setting, we randomize eight vortices and six obstacles to create challenging layouts under severe noise. We train the models for 1.5M time steps over nine seeds and evaluate every 10,000 steps on 15 pre-randomized environments, resetting the environmental layout each episode. For the UAV setting, we keep six obstacles and follow the original experimental protocol with 200,000 training frames.

Noise and Grouping Settings. The critical challenge lies in maintaining the robustness of DistRL under noisy observation. In order to narrow the gap between simulation and real sensing, we ground our noise modeling in prior physical studies [Koenig and Howard, 2004; Laconte *et al.*, 2021] and superimpose ubiquitous non-synthetic sensor noise processes, specifically Gaussian and Poisson noise. To assess DRIQN's multi-noise robustness, we incorporate extra salt-and-pepper noise (simulating extreme value fluctuations) and occlusion noise (emulating partial sensor failures) in Section 4.4. The noise intensity is treated as a tunable hyperparameter and is empirically set to $\{0.6, 0.4, 0.2\}$ to systematically assess performance degradation under different perturbation levels; for the cluttered setting, we fix the intensity at 0.2. The sensor value range and noise intensity jointly determine the resulting distributions. As our focus is enhancing robustness in online DistRL training via DRO, we adopt a simple, practice-aligned grouping scheme based on collection periods with known noise types rather than clustering: data collected within the same period forms one subgroup, reflecting that noise conditions naturally shift across time and regions. This DRO framework is agnostic to the specific noise law, enabling extensions to other settings, including safety analysis under adversarial perturbations.

Baselines. We compare DRIQN with the following representative methods in autonomous navigation: (1) Classical observation-based planning methods: Artificial Potential Field (APF) method [Fan *et al.*, 2020] and Bug Algorithm (BA) method [Lumelsky and Skewis, 1988]; (2) Standard Q-learning method in DRL: DQN [Mnih *et al.*, 2015]; (3) Risk-sensitive method in DistRL: IQN [Dabney *et al.*, 2018].

Hyperparameter Settings. For neural networks with three or more layers, we implement gradient substitution exclusively in the final layer instead of the whole layers, inspired by established computer vision research on gradient manipulation [Shen *et al.*, 2022]. The comparative analysis is detailed in Section 4.5. Additionally, to adapt the gradients computed by the DRO to the training of all layers except the final one, we adapt the learning rate decay of these layers starting from 0.0001 to 0.000001 and appropriately shrink the calculated gradients. The implementation builds upon previous work mentioned in Section 4.1, with other hyperparameters configured in accordance with their respective network settings. The experiments were completed in parallel using six A6000 GPUs and AMD EPYC 7542 CPU.

4.2 Metrics and Strategy

Metrics. To evaluate the performance of our approach, we employed six metrics encompassing performance, safety, and efficiency: success rate, collision rate, timeout rate, final reward, average time consumed, and average energy consumed. Average time and energy are calculated based solely on data from successful episodes. Given the physical interpretation of these metrics, path smoothness can be indirectly quantified via the resulting time and energy costs.

Strategy. We compare a standard greedy strategy that selects the action with the highest expected return with the adaptive

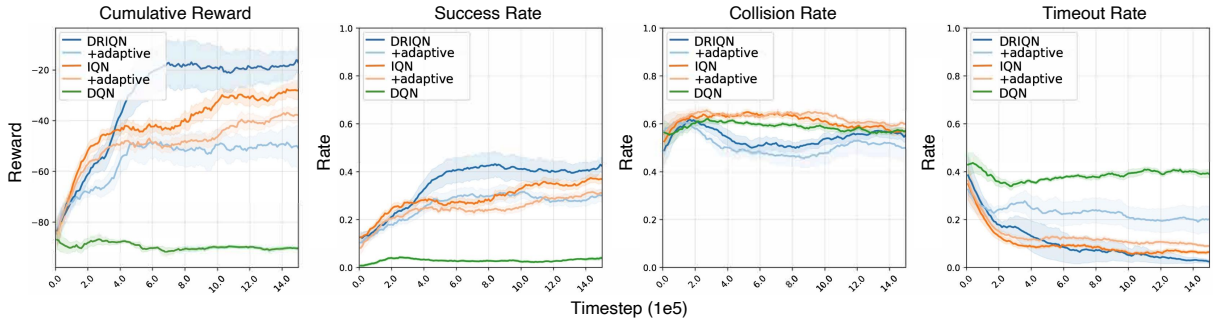

Figure 3: Online evaluations of the marine environment compare three learning-based models and two distributional RL strategies. Performance, safety, and convergence are assessed via three key metrics with cumulative reward curves. Implementation details appear in Section 4.1. Due to space limit, the other two figures beyond noise level 0.6 are in Supplementary Material, Section B.

Method	Strategy	SR ↑			CR ↓			TR ↓			FCR ↑			AT ↓			AE ↓		
		0.6	**0.4**	**0.2**	**0.6**	**0.4**	**0.2**	**0.6**	**0.4**	**0.2**	**0.6**	**0.4**	**0.2**	**0.6**	**0.4**	**0.2**	**0.6**	**0.4**	**0.2**
APF	–	0.11	0.25	0.03	0.81	0.55	0.53	0.08	0.20	0.44	–	–	–	206.68	349.32	547.89	320.96	654.48	1071.70
BA	–	0.22	0.11	0.26	0.71	0.46	**0.33**	0.07	0.43	0.41	–	–	–	185.76	482.98	462.97	251.11	691.93	691.57
DQN	Greedy	0.04	0.03	0.02	0.57	0.57	0.57	0.39	0.40	0.41	-90.30	-90.11	-89.54	541.27	550.35	560.52	858.39	779.78	832.67
IQN	Greedy	0.37	0.50	0.61	0.57	0.45	0.37	0.06	**0.04**	0.02	-28.16	-14.87	13.47	144.89	123.75	61.25	196.73	181.63	92.18
	Adaptive	0.31	0.36	0.46	0.60	0.51	0.38	0.09	0.13	0.16	-37.93	-38.97	-48.66	185.75	230.00	274.27	247.34	296.00	358.99
DRIQN	Greedy	**0.42**	**0.55**	**0.63**	0.55	**0.42**	0.37	**0.02**	**0.04**	**0.01**	**-17.11**	**3.14**	**26.87**	**93.51**	**90.11**	**31.82**	**141.67**	**114.15**	**53.61**
	Adaptive	0.30	0.39	0.33	**0.50**	0.46	0.49	0.20	0.15	0.18	-50.84	-36.16	-45.90	283.85	226.07	272.25	392.90	257.17	401.26
Improvement (%)		**13.51**	**10.00**	**3.28**	**12.28**	**6.67**	-12.12	**66.67**	–	**50.00**	**39.24**	**121.12**	**99.48**	**35.46**	**27.18**	**48.05**	**27.99**	**37.15**	**41.84**

Table 1: **Overall performance comparison** under three different noise levels (0.6, 0.4, and 0.2) related to the variances of the noise distributions. The best results are highlighted in **bold** and the second best results are highlighted with an underline. "Improvement" denotes the percentage increase (positive metrics) or decrease (negative metrics) over the second best models. For brevity, we use corresponding abbreviations for the metrics in this and all subsequent tables: success rate (SR), collision rate (CR), timeout rate (TR), final cumulative reward (FCR), average time consumed (AT), and average energy consumed (AE).

IQN strategy of [Lin *et al.*, 2023], which adjusts the CVaR threshold based on real-time risk estimates (proxied by distance to the nearest obstacle).

4.3 Performance Comparison

Online evaluation results in Figure 3 (reported for the marine setting, which adopts the more challenging training protocol) are smoothed using an Exponential Moving Average to highlight learning dynamics; shaded regions denote the standard deviation over nine seeds under high noise intensity. Through multiple trials, the shaded regions demonstrate the significance of DRIQN's performance. DRIQN adds an overhead of only 0.018ms per update and outperforms the runner-up.

For USV navigation, the qualitative trajectory results of RL agents are visualized in Figure 4. Key performance illustrated in Table 1 are further summarized as follows. The improvements are statistically significant (e.g., task-critical metrics Success: $p = 0.0093$; Collision: $p = 0.0380$).

- *DRIQN vs. Baseline Methods*: As a DistRL approach enhanced by DRO-based gradient substitution, DRIQN achieves superior success / collision rates and impressive efficiency compared with other traditional physical methods and RL algorithms across noise intensities, particularly under high noise variance.
- *DistRL vs. Standard RL*: DistRL methods consistently outperform standard deep Q-learning. DQN exhibits high timeout rates and resource consumption due to passive collision avoidance, stemming from its reliance on expected returns that are less informative in noisy settings. In contrast, DistRL explicitly models return distributions, yielding stronger noise robustness.
- *Greedy vs. Adaptive*: Both IQN and DRIQN achieve higher success rates with greedy action selection. Adaptive strategies degrade under noise, as perturbed value distributions undermine their decision rule.
- *DRIQN Efficiency*: Under stringent noise-corrupted conditions, greedy DRIQN maintains the lowest timeout rate while reducing energy and time consumption, optimally balancing safety and efficiency.

The additional drone navigation experiments in Table 2 further validate DRIQN's superiority under observational noise, supporting its generalization across robot classes and navigation environments. Qualitative trajectories from IQN and DRIQN are shown in Figure 5.

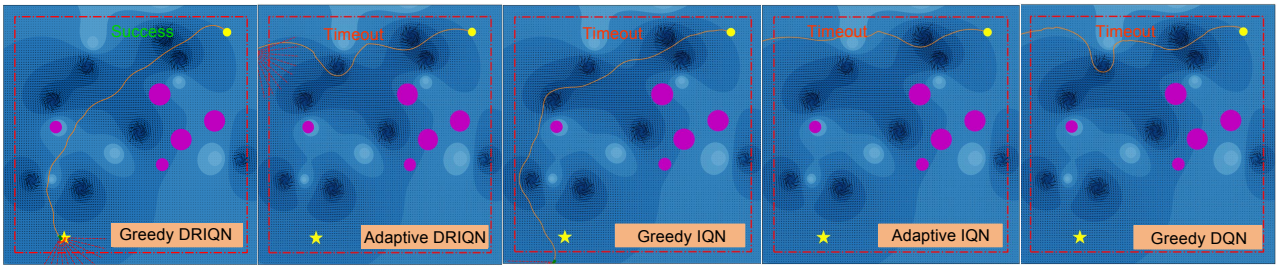

Figure 4: **Qualitative trajectory results** (noise variance level = 0.6). Yellow circle: start position, yellow star: the goal, Magenta circles: static obstacles. The greedy approach avoids risk-sensitive adjustments triggered by noisy observations, unlike the adaptive strategy. Aligning with Table 1, the adaptive strategy exhibits higher timeout rates and frequent deviations from complex conditions to mitigate risk.

Method (Greedy)	SR ↑			FCR ↑		
	0.6	0.4	0.2	0.6	0.4	0.2
IQN	0.20	0.35	0.27	-16.63	-3.20	-7.85
DRIQN	**0.43**	**0.41**	**0.46**	**0.85**	**0.50**	**5.06**

Table 2: Drone navigation performance under all noise levels, evaluated with greedy action selection on 100 randomly generated noisy environments with six obstacles each.

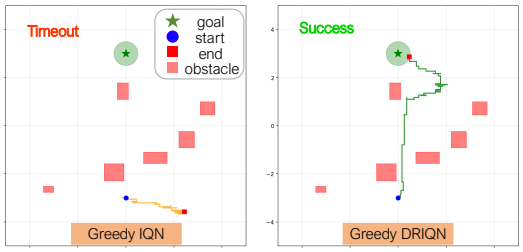

Figure 5: Qualitative drone trajectories generated by IQN (left) and DRIQN (right) under strong noise level 0.6.

Model	SR ↑	CR ↓	TR ↓	FCR ↑	AT ↓	AE ↓
APF	0.12	0.80	0.08	–	211.83	351.38
BA	0.14	0.82	0.04	–	134.65	198.20
DQN	0.11	0.50	0.39	-80.02	507.04	798.70
IQN	0.54	0.44	0.02	-3.32	75.27	108.54
DRIQN	**0.58**	**0.41**	**0.01**	**12.45**	**45.10**	**73.30**

Table 3: Robustness comparison under the multi-noise setting, where natural observation noise is combined with salt-and-pepper noise and occlusion noise.

Model	Strategy	SR ↑	CR ↓	AT ↓
DRIQN	Greedy	**0.42**	0.55	**93.51**
	Adaptive	0.30	**0.50**	283.85
DRIQN-W	Greedy	0.22	0.54	357.87
	Adaptive	0.20	0.55	379.50

Table 4: Architectural comparison of "gradient substitution" under high noise variance: Whole-network substitution (DRIQN-W) versus final-layer substitution (DRIQN).

4.4 The Analysis of "Noise Types"

To rigorously evaluate DRIQN's safety in multi-noise conditions, we augment natural noise conditions with salt-and-pepper noise (simulating sensor malfunction) and occlusion noise (emulating data loss scenarios) for faulty sensors, setting a fractionally lower noise intensity across all sufficient perturbations for intensity-diversity balance. The strategy is rigged to greedy rather than adaptive for more competitive performance in this task, as empirically demonstrated earlier. Evidenced in Table 3, DRIQN's optimization across heterogeneous noise subgroups demonstrates superior handling of complex perceptual patterns and observational robustness over baselines. We also try to replace the original grouping with success and trajectory length grouping. New partitioning drops the Success by 73% and increases the Collision by 4% at 0.6 intensity, justifying rationality of the original noise type based grouping.

4.5 The Analysis of "Gradient Substitution"

To examine our last-layer gradient substitution setting for DRIQN, we introduce the variant DRIQN-W featuring whole-network (4-layer) gradient replacement for comparison. As evidenced in Table 4, last-layer substitution significantly outperforms whole-network gradient replacement in both success rate and computational efficiency while ensuring better safety. This advantage stems from retaining deep networks' representational capacity while improving robustness via streamlined optimization, even under extreme noise. Learning curves (Figure S2 in Supplementary Material) show that last-layer substitution converges faster and more stably than full-network modification.

5 Conclusion

In this paper, we propose DRIQN, a robust and efficient DistRL policy for path planning under observation noise. DRIQN combines flexible gradient substitution with DRO-guided gradient computation to improve implicit-quantile DistRL under heterogeneous noise. Experiments show consistent gains over strong baselines and validate robustness across tasks and vehicle types. We further find that greedy action selection is more reliable than distance-based adaptation under noisy observations, highlighting the fragility of handcrafted physical cues. Future work will extend DRIQN to continuous control and real-world deployment.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (Grant No.2023YFF0725001), in part by the National Natural Science Foundation of China (Grant No.92370204), in part by the guangdong Basic and Applied Basic Research Foundation (Grant No.2023B1515120057), in part by the Key-Area Special Project of Guangdong Provincial Ordinary Universities(2024ZDZX1007), in part by the Department of Science and Technology of Guangdong Province (2023CX10X079), the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2023A03J0008), and Education Bureau Guangzhou Municipality.

References

[Asfihani *et al.*, 2025] Tahiyatul Asfihani, Ahmad Maulana Syafi'i, and Agus Hasan. Fault-tolerant model predictive control for unmanned surface vehicles. *Systems Science & Control Engineering*, 13(1):2469598, 2025.

[Dabney *et al.*, 2018] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

[Dawood *et al.*, 2024] Murad Dawood, Ahmed Shokry, and Maren Bennewitz. A dynamic safety shield for safe and efficient reinforcement learning of navigation tasks. *ArXiv*, abs/2412.04153, 2024.

[Elmakis and Degani, 2023] Oren Elmakis and Amir Degani. Usv port oil spill cleanup using hybrid multi-destination rl-cpp. *IEEE Access*, 11:122722–122735, 2023.

[Fan *et al.*, 2020] Xiaojing Fan, Yinjing Guo, Hui Liu, Bowen Wei, and Wenhong Lyu. Improved artificial potential field method applied for auv path planning. *Mathematical Problems in Engineering*, 2020.

[Fernandes *et al.*, 2025] João FP Fernandes, Mário Assunção, Daniel Serrano, Pedro Afonso, Pedro Pinheiro, Hugo Marques, José Neves, Pedro Teodoro, Ricardo Póvoa, Rosa Marat-Mendes, et al. A smart energy management system for surface unmanned vehicles for border surveillance missions. *Scientific Reports*, 15(1):23684, 2025.

[Han *et al.*, 2016] Jungwook Han, Jeonghong Park, Jinwhan Kim, and Nam-sun Son. Gps-less coastal navigation using marine radar for usv operation. *Ifac-papersonline*, 49(23):598–603, 2016.

[Hazan *et al.*, 2006] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69:169–192, 2006.

[He *et al.*, 2025] Yiting He, Zhishuai Liu, Weixin Wang, and Pan Xu. Sample complexity of distributionally robust off-dynamics reinforcement learning with online interaction. In *Forty-second International Conference on Machine Learning*, 2025.

[Hong *et al.*, 2024] Liang Hong, Haitao Liu, Quanshun Yang, and Jiaxuan Yao. Model predictive attitude control of unmanned surface vehicle based on short-time wave prediction. *Ocean Engineering*, 314:119727, 2024.

[Hossain *et al.*, 2024] Jumman Hossain, Abu-Zaher Faridee, Derrik Asher, Jade Freeman, Theron Trout, Timothy Gregory, and Nirmalya Roy. Quasinav: Asymmetric cost-aware navigation planning with constrained quasimetric reinforcement learning. *arXiv preprint arXiv:2410.16666*, 2024.

[Huber, 1992] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[Khatib, 1985] Oussama Khatib. Real-time obstacle avoidance for manipulators and mobile robots. *The International Journal of Robotics Research*, 5:90 – 98, 1985.

[Koenig and Howard, 2004] Nathan Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2149–2154, 2004.

[Laconte *et al.*, 2021] Johann Laconte, Elie Randriamiarintsoa, Abderrahim Kasmi, F. Pomerleau, Roland Chapuis, Christophe Debain, and Romuald Aufrère. Dynamic lambda-field: A counterpart of the bayesian occupancy grid for risk assessment in dynamic environments. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4846–4853, 2021.

[Leung *et al.*, 2025] Cheuk Hang Leung, Yiyan Huang, Yijun Li, and Qi Wu. Distributionally robust policy evaluation and learning for continuous treatment with observational data. *ArXiv*, abs/2501.10693, 2025.

[Lin *et al.*, 2023] Xi Lin, John McConnell, and Brendan Englot. Robust unmanned surface vehicle navigation with distributional reinforcement learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6185–6191. IEEE, 2023.

[Liu *et al.*, 2022] Cheng Liu, Erik-Jan van Kampen, and Guido CHE De Croon. Adaptive risk-tendency: Nano drone navigation in cluttered environments with distributional reinforcement learning. *arXiv preprint arXiv:2203.14749*, 2022.

[Liu *et al.*, 2023] Cheng Liu, Erik-Jan van Kampen, and Guido CHE De Croon. Adaptive risk-tendency: Nano drone navigation in cluttered environments with distributional reinforcement learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7198–7204. IEEE, 2023.

[Liu *et al.*, 2025] Guangyi Liu, Suzan Iloglu, Michael Caldara, Joseph W Durham, and Michael M Zavlanos. Distributionally robust multi-agent reinforcement learning for dynamic chute mapping. *arXiv preprint arXiv:2503.09755*, 2025.

[Lolla *et al.*, 2015] Tapovan Lolla, Patrick J Haley Jr, and Pierre FJ Lermusiaux. Path planning in multi-scale ocean

flows: Coordination and dynamic obstacles. *Ocean Modelling*, 94:46–66, 2015.

[Lumelsky and Skewis, 1988] Vladimir J. Lumelsky and Tim Skewis. A paradigm for incorporating vision in the robot navigation function. *Proceedings. 1988 IEEE International Conference on Robotics and Automation*, pages 734–739 vol.2, 1988.

[Lumelsky and Stepanov, 1986] Vladimir J. Lumelsky and Alexander A. Stepanov. Dynamic path planning for a mobile automaton with limited information on the environment. *IEEE Transactions on Automatic Control*, 31:1058–1063, 1986.

[Lütjens *et al.*, 2018] Björn Lütjens, Michael Everett, and Jonathan P. How. Safe reinforcement learning with model uncertainty estimates. *2019 International Conference on Robotics and Automation (ICRA)*, pages 8662–8668, 2018.

[Maharjan *et al.*, 2022] Nisha Maharjan, Hiroyuki Miyazaki, Bipun Man Pati, Matthew N Dailey, Sangam Shrestha, and Tai Nakamura. Detection of river plastic using uav sensor data and deep learning. *Remote Sensing*, 14(13):3049, 2022.

[McGuire *et al.*, 2018] Kimberly N. McGuire, Guido C.H.E. de Croon, and Karl Tuyls. A comparative study of bug algorithms for robot navigation. *Robotics Auton. Syst.*, 121, 2018.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Kirkeby Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[Nietert *et al.*, 2023] Sloan Nietert, Ziv Goldfeld, and Soroosh Shafiee. Outlier-robust wasserstein dro. *Advances in Neural Information Processing Systems*, 36:62792–62820, 2023.

[Qi *et al.*, 2020] Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. An online method for a class of distributionally robust optimization with non-convex objectives. In *Neural Information Processing Systems*, 2020.

[Ramesh *et al.*, 2024] Shyam Sundhar Ramesh, Pier Giuseppe Sessa, Yifan Hu, Andreas Krause, and Ilija Bogunovic. Distributionally robust model-based reinforcement learning with large state spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2024.

[Shen *et al.*, 2022] Xuli Shen, Xiaomei Wang, Qing Xu, Weifeng Ge, and Xiangyang Xue. Towards scalable and fast distributionally robust optimization for data-driven deep learning. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 448–457. IEEE, 2022.

[Shi and Chi, 2024] Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024.

[Shi *et al.*, 2024a] Jiyuan Shi, Chenjia Bai, Haoran He, Lei Han, Dong Wang, Bin Zhao, Mingguo Zhao, Xiu Li, and Xuelong Li. Robust quadrupedal locomotion via risk-averse policy learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11459–11466. IEEE, 2024.

[Shi *et al.*, 2024b] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36, 2024.

[Staib and Jegelka, 2019] Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[Sun *et al.*, 2019] Jiubo Sun, Guoliang Liu, Guohui Tian, and Jianhua Zhang. Smart obstacle avoidance using a danger index for a dynamic environment. *Applied Sciences*, 2019.

[Tobin *et al.*, 2017] Joshua Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017.

[Tomilin *et al.*, 2025] T. Tomilin, M. Fang, and M. Pechenizkiy. Hasard: A benchmark for vision-based safe reinforcement learning in embodied agents. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Vieira *et al.*, 2020] Manuel Vieira, M Clara P Amorim, Andreas Sundelöf, Nuno Prista, and Paulo J Fonseca. Underwater noise recognition of marine vessels passages: Two case studies using hidden markov models. *ICES Journal of Marine Science*, 77(6):2157–2170, 2020.

[Wang *et al.*, 2023a] Jincheng Wang, Lei Xia, Lei Peng, Huiyun Li, and Yunduan Cui. Efficient uncertainty propagation in model-based reinforcement learning unmanned surface vehicle using unscented kalman filter. *Drones*, 7(4):228, 2023.

[Wang *et al.*, 2023b] Yuanda Wang, Wenzhang Liu, Jian Liu, and Changyin Sun. Cooperative usv–uav marine search and rescue with visual navigation and reinforcement learning-based control. *ISA transactions*, 137:222–235, 2023.

[Wang *et al.*, 2024] Junqiao Wang, Zhongliang Yu, Dong Zhou, Jiaqi Shi, and Runran Deng. Vision-based deep reinforcement learning of uav autonomous navigation using privileged information. *arXiv preprint arXiv:2412.06313*, 2024.

[Zhang *et al.*, 2025] Heng Zhang, Guoxiang Zhao, and Xiaoqiang Ren. Terl: Large-scale multi-target encirclement using transformer-enhanced reinforcement learning. *arXiv preprint arXiv:2503.12395*, 2025.