Expert Systems With Applications 308 (2026) 131118

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Human-in-the-loop reinforcement learning with risk-aware intervention and imitation

Yaqing Zhou [a], Yun-Bo Zhao [a,b,*], Chenwei Xu [a], Chen Ouyang [a], Pengfei Li [a]

[a] Department of Automation, University of Science and Technology of China, Hefei, China
[b] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China

ARTICLE INFO

Keywords:
Human-in-the-loop
Reinforcement learning
Human intervention
Imitation learning

ABSTRACT

Human-in-the-loop reinforcement learning (HiL-RL) improves policy safety and learning efficiency by incorporating real-time human interventions and demonstration data. However, existing HiL-RL methods often suffer from inaccurate intervention timing and inefficient use of demonstration data. To address these issues, we propose a novel framework called HiRIL (Human-in-the-loop Risk-aware Imitation-enhanced Learning), which establishes a closed-loop learning mechanism that integrates risk-aware intervention triggering and imitation-based policy optimization under a dual-mode uncertainty metric. At the core of HiRIL is the Bayesian Implicit Quantile Network (BIQN), which captures both epistemic and aleatoric uncertainty through Bayesian weight sampling and quantile-based return modeling. These uncertainties are combined to generate risk scores for state-action pairs, guiding when to trigger human intervention. To better utilize intervention data, HiRIL introduces a prioritized experience replay mechanism based on risk difference, which emphasizes human interventions that significantly reduce risk. During policy optimization, a local imitation loss is applied to clone human actions at intervention points, enabling risk-guided joint optimization. We conduct extensive experiments on the CARLA end-to-end autonomous driving benchmark. Results show that HiRIL consistently outperforms baselines across multiple metrics and maintains strong robustness under perturbations and non-stationary human intervention.

1. Introduction

Human-in-the-loop reinforcement learning (HiL-RL) has increasingly become a key technology for enhancing the applicability of reinforcement learning in safety-critical scenarios in recent years (Retzlaff et al., 2024; Yu & Chang, 2022; Zhang et al., 2021). Unlike traditional deep reinforcement learning, which relies solely on autonomous exploration, HiL-RL allows humans to intervene during training when the agent engages in high-risk behaviors, providing high-value demonstration data through takeover actions (Wu et al., 2022a,b). In complex tasks with sparse rewards or high risks, this interactive paradigm not only reduces potential dangers during the agent's exploration process but also leverages human prior knowledge to improve learning efficiency and guide the learning direction. It offers a more reliable training framework for domains such as autonomous driving, robotic manipulation, and large-scale model training and alignment (Lanzaro & Sayed, 2024; Liu, 2025; Tan et al., 2025).

Despite the significant potential of HiL-RL, existing research still faces two core limitations. On one hand, regarding the intervention triggering mechanism, early methods rely on continuous human supervision to prevent unsafe actions through safety interruptions (Amodei et al., 2016) or manual takeovers (Saunders et al., 2017). However, these approaches are labor-intensive and difficult to scale. To reduce the human workload, some studies have attempted to enable agents to autonomously identify dangerous states by employing external risk predictors (Xie et al., 2022a), artificial potential fields (Huang et al., 2024), or single-type uncertainty measures (Singi et al., 2024) to trigger interventions. Nevertheless, these methods depend on additional modules or adopt single-mode uncertainty modeling, making it difficult to accurately assess the risk level of the policy. This often leads to either excessive or insufficient human intervention. On the other hand, in terms of utilizing intervention data, although imitation learning methods can leverage demonstration data from interventions to accelerate training (Nair et al., 2018), large volumes of autonomously collected experience tend to dilute the quality of limited demonstrations, making it difficult for them to have a lasting impact on policy updates. Even with dual-buffer experience replay mechanisms (Liu et al., 2025), the variation in demonstration quality is often overlooked, resulting in high-value

[*] Corresponding author.
E-mail address: ybzhao@ustc.edu.cn (Y.-B. Zhao).

https://doi.org/10.1016/j.eswa.2026.131118
Received 15 October 2025; Received in revised form 10 December 2025; Accepted 3 January 2026
Available online 5 January 2026

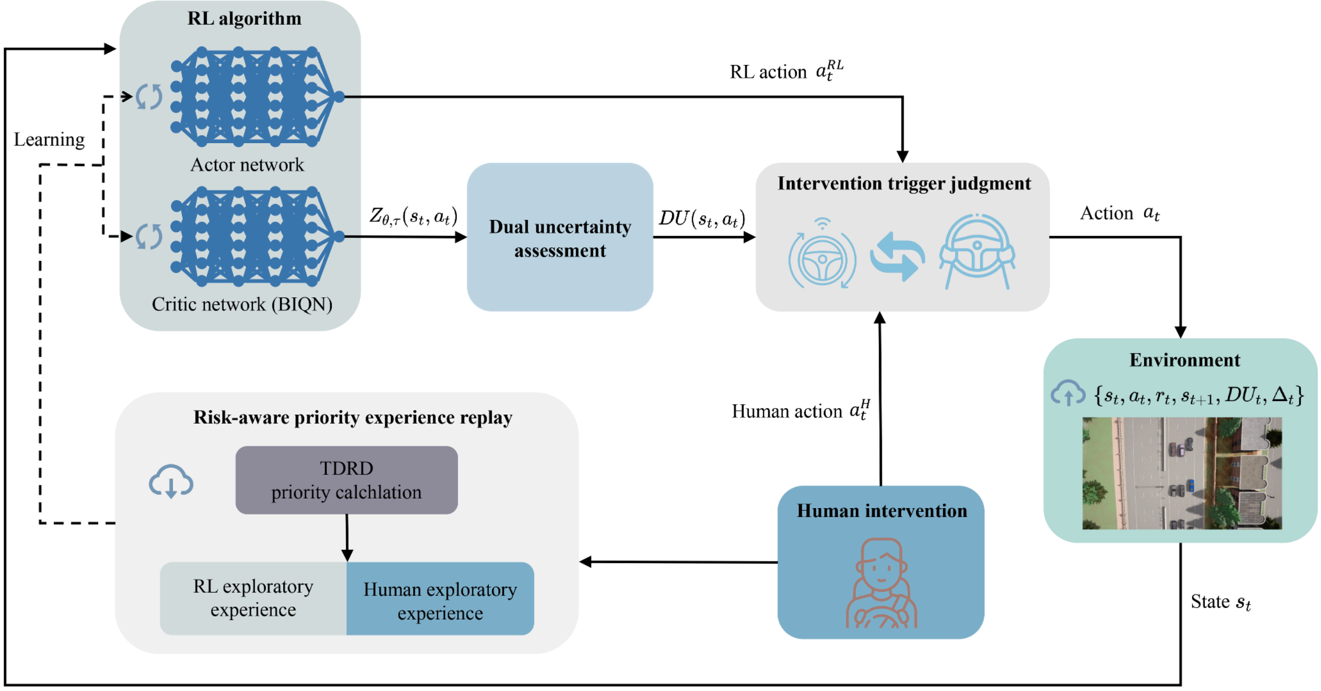
RL algorithm
Actor network
Critic network (BIQN)
Learning
RL action a_t^{RL}
$Z_{\theta,\tau}(s_t,a_t)$
Dual uncertainty assessment
$DU(s_t,a_t)$
Intervention trigger judgment
Action a_t
Environment
$\{s_t,a_t,r_t,s_{t+1},DU_t,\Delta_t\}$
Human action a_t^H
Risk-aware priority experience replay
TDRD priority calchlation
RL exploratory experience
Human exploratory experience
Human intervention
State s_t

Fig. 1. Overall architecture of the HiRIL framework. The environment state is first fed into the actor network to generate an action, while the critic network, built upon BIQN, outputs the quantile return distribution for each state-action pair to perform dual uncertainty assessment. If the estimated uncertainty exceeds a predefined threshold, a human intervention is triggered and the human-provided action is executed; otherwise, the RL action is used. The environment then returns transition information, rewards, uncertainty values, and risk labels, which are recorded together with the state and action. These transitions are categorized into exploratory and intervention experiences and stored in a risk-difference-based prioritized replay buffer. This enables subsequent policy updates guided by risk-aware imitation learning.

intervention data not being fully utilized. In summary, current HiL-RL methods still suffer from two major bottlenecks: inaccurate intervention timing and inefficient utilization of intervention demonstrations.

To address the above challenges, this paper proposes a Human-in-the-loop Risk-aware Imitation-enhanced Learning framework (HiRIL). This framework generates risk assessments through end-to-end uncertainty modeling, enabling precise intervention triggering, and fully leverages human intervention data via risk-sensitive experience replay and imitation-enhanced learning. The overall architecture of the proposed method is illustrated in Fig. 1. Specifically, we introduce the Bayesian Implicit Quantile Network (BIQN) as a risk scorer. BIQN estimates epistemic uncertainty through Bayesian sampling and captures aleatoric uncertainty via quantile-based return modeling, thereby generating precise risk assessments for each state-action pair that incorporate both types of uncertainty. Based on these risk scores, HiRIL proactively triggers human takeovers in high-risk states and applies a one-time penalty to guide the policy away from dangerous regions. During experience replay, we design a risk-difference-based prioritized replay mechanism to better utilize key demonstration segments where human actions significantly reduce risk. In the policy optimization phase, we incorporate a behavior cloning objective to improve learning efficiency.

This paper makes the following contributions:

1. **A comprehensive risk-aware human-in-the-loop framework.** We propose **HiRIL**, this framework integrates risk-triggered human intervention, risk-difference-based prioritized replay, and imitation-enhanced policy updates into a closed-loop training process, significantly improving both safety and learning efficiency.
2. **A Bayesian distributional risk estimator.** We propose **BIQN**, which provides distributional representations for both epistemic and aleatoric uncertainties, and offers an unbiased estimation of the second-order moment of their joint distribution. This estimate is used as a risk metric to enable more precise triggering of human interventions.
3. **Systematic empirical validation in continuous control.** On the CARLA end-to-end autonomous driving benchmark, HiRIL demonstrates comprehensive advantages over existing baselines across all aspects, including both training performance and testing robustness, showcasing its exceptional overall capabilities.

The remainder of the paper is organized as follows: Section 2 reviews the related work. Section 3 provides the preliminaries. Section 4 introduces the BIQN architecture and defines the risk estimation method. Section 5 presents the proposed HiRIL approach. Section 6 describes the experimental setup and analyzes the results. Section 7 concludes the paper.

2. Related work

Human Intervention. Incorporating human intervention into reinforcement learning has been widely recognized as an effective approach to improving agent safety and training reliability (Tan et al., 2025; Xie et al., 2022b; Xu et al., 2022). Early methods largely relied on continuous human supervision. For example, the safety interruption mechanism proposed in Amodei et al. (2016) allows real-time termination of dangerous behaviors, while Saunders et al. (2017) trains a supervisory model through human interception of unsafe actions to reduce the frequency of future interventions. However, such methods require expert involvement throughout the training process, resulting in high labor costs. To alleviate this issue, some studies have explored mechanisms that allow agents to actively request interventions. For instance, Mandel et al. (2017) employs neural networks to identify high-risk states, Xie et al. (2022a) detects irreversible states to request assistance, and Huang et al. (2024) uses artificial potential fields to determine the timing of interventions. While these approaches reduce human workload, they rely on external modules and suffer from limited generalization capabilities. In recent years, some works have shifted toward leveraging the agent's internal uncertainty estimates to develop end-to-end risk-aware intervention strategies without external structures. For example, Singi et al. (2024) focuses on epistemic uncertainty, and Silva et al. (2020) targets aleatoric uncertainty. In contrast, this paper proposes a more refined dual-mode uncertainty modeling approach that combines Bayesian sampling with quantile-based return modeling, enabling more accurate risk quantification and more reliable intervention triggering.

Imitation Learning. Imitation learning improves the learning efficiency and convergence speed of agents in environments with sparse rewards or high complexity by leveraging human intervention behaviors as demonstration data (Celemin et al., 2022; Hua et al., 2021; Zare et al., 2024). Representative methods such as DQfD (Hester et al., 2018) and GAIL (Ho & Ermon, 2016) can significantly boost performance in the early stages of training. However, these approaches mainly rely on action substitution or supervised pretraining and do not involve direct optimization of the policy structure itself. In recent years, some studies have explored deeper integration of human expertise by introducing behavior cloning objectives (Fujimoto & Gu, 2021) or modifying the policy function (Nair et al., 2018). Nevertheless, limited human demonstrations are still prone to being diluted by large volumes of autonomously collected experience. To mitigate this issue, several works have proposed dual experience replay mechanisms, storing human demonstrations and agent-generated experiences in separate buffers (Liu et al., 2025; Wang et al., 2018). While this improves the utilization efficiency of demonstration data, it fails to account for the inherent variability in demonstration quality. Building upon the prioritized experience replay mechanism (Schaul et al., 2015), this study introduces a risk-difference term and combines penalty-based reward shaping with imitation-enhanced joint policy optimization, enabling more effective exploitation of high-value intervention data and enhancing overall policy performance.

Uncertainty in Deep Reinforcement Learning. Uncertainty modeling in deep reinforcement learning primarily aims to optimize the exploration-exploitation trade-off and improve training efficiency. Existing research typically categorizes uncertainty into two types (Gawlikowski et al., 2023): epistemic uncertainty, which arises from the agent's lack of knowledge about unknown states, and aleatoric uncertainty, which originates from the inherent stochasticity of the environment. Epistemic uncertainty is commonly modeled using ensemble methods (e.g., Bootstrapped DQN Osband et al., 2016) and Bayesian approaches (van der Vaart et al., 2024). While ensemble methods heuristically estimate uncertainty, they often lack proper probabilistic calibration. In contrast, Bayesian methods introduce distributions over model parameters and offer stronger theoretical guarantees. Notably, Dropout (Hiraoka et al., 2021) can be interpreted as a form of Bayesian approximation. Due to its lower computational cost, it is widely used in practice. Aleatoric uncertainty is typically modeled using distributional reinforcement learning, such as QR-DQN (Dabney et al., 2018b) and IQN (Dabney et al., 2018a). IQN employs implicit quantile regression, allowing for more flexible and accurate modeling of return distributions. Modeling a single type of uncertainty is often insufficient (Lockwood & Si, 2022); hence, joint modeling of both types has gained increasing attention. EQN (Hoel et al., 2023), for example, attempts to combine ensembles with distributional regression to capture both epistemic and aleatoric uncertainty. However, this approach yields biased estimates of variance (Clements et al., 2019). To address this limitation, we propose the BIQN, which integrates Bayesian inference with quantile-based modeling in a principled manner. BIQN provides a full distributional model that simultaneously captures both types of uncertainty, offering stronger representation capabilities. Furthermore, this modeling approach enables unbiased variance estimation via Monte Carlo sampling.

3. Preliminaries

In this section, we first introduce the notation and concepts of HiL-RL, followed by the quantile-based methods for modeling aleatoric uncertainty and the Bayesian approaches for modeling epistemic uncertainty. The three components presented in this section collectively form the foundation of the proposed HiRIL method.

3.1. Notation

We model the interaction between the RL agent and the environment as a Markov Decision Process (MDP) $M = (S, A, R, P, \mu_0, \gamma)$, where S denotes the state space, A denotes the action space, R is the reward function, P is the transition probability, μ_0 represents the initial state distribution, and γ is the discount factor.In this paper, we adopt a human-in-the-loop learning framework, where humans can proactively intervene to control agent.

Given the current state $s_t \in S$, the agent samples an action $a_t^{RL} \in A$ from the policy $\pi_\phi(s_t)$, while a human can override this action with a human action $a_t^H \in A$. Therefore, the executed action is defined as:

$$a_t = \Delta_t a_t^H + (1 - \Delta_t) a_t^{RL},$$

where Δ_t is a binary indicator function of human intervention.

3.2. IQN: Implicit quantile networks

Unlike traditional Q-learning, IQN (Dabney et al., 2018a) belong to a class of distributional RL methods. They focus on the inherent randomness of returns within the RL framework and aim to model the distribution of returns. IQN models return values through implicit quantiles:

$$Z_\tau := F_Z^{-1}(\tau) = f_\tau(s, a). \tag{1}$$

where Z_τ represents the return's quantile function evaluated at $\tau \sim \mathcal{U}(0, 1)$.

The training objective of IQN is to reparameterize samples from the base distribution to match the corresponding quantiles of the target distribution. For two quantile samples $\tau, \tau' \sim \mathcal{U}(0, 1)$, the sampled temporal difference (TD) error at step t is:

$$\delta_t^{\tau,\tau'} = r_t + \gamma Z_{\tau'}\left(s_{t+1}, \pi^*(s_{t+1}); \theta^-\right) - Z_\tau(s_t, a_t; \theta) \tag{2}$$

where $\pi^*(s) = \arg\max_a Q(s, a)$, the sample-based $Q^\pi(s_t, a_t)$ estimation is calculated by drawing K_τ samples from $\tau \sim \mathcal{U}(0, 1)$:

$$\tilde{Q}^\pi(s, a) = \frac{1}{K_\tau} \sum_{k=1}^{K_\tau} Z_{\tau_k}(s, a; \theta) \tag{3}$$

The loss function of IQN is defined as:

$$\mathcal{L}_{IQN}(\theta) = \mathbb{E}_{s_t \sim D}\left[\frac{1}{N} \sum_{i,j} \rho_\kappa\left(\delta_t^{\tau_i,\tau'_j}\right)\right] \tag{4}$$

where ρ_κ is the Huber quantile regression loss, and D is the replay buffer that stores training data.

3.3. BNN: Bayesian neural network

Bayesian Neural Networks (BNN) (Goan & Fookes, 2020) treat weights and biases as random variables and sample the network's weights from a posterior distribution $p(\theta|\mathcal{D})$, where $\mathcal{D} = (X, Y)$ represents the experiences collected by the agent, $X = (x_i)_{i=1}^M$ is the input set, and $Y = (y_i)_{i=1}^M$ is the target label set. Since $p(\theta|D)$ is difficult to compute, a variational distribution $q(\theta)$ is typically sampled to approximate it, i.e. maximizing the Evidence Lower Bound (ELBO):

$$\begin{aligned} ELBO &= \int q(\theta) \log p(Y \mid X, \theta) d\theta - KL(q(\theta)|p(\theta)) \\ &= \left(\sum_{i=1}^{M} \int q(\theta) \log p\left(y_i \mid x_i, \theta\right) d\theta\right) - KL(q(\theta)|p(\theta)) \end{aligned} \tag{5}$$

4. Distributional risk estimation via BIQN

This section introduces the BIQN model, which integrates IQN with BNN to construct a unified distributional representation of both epistemic and aleatoric uncertainties. Based on the resulting joint distribution, we further provide an unbiased estimation of its second-order moment, thereby offering a quantitative metric for risk assessment to trigger human interventions and effectively leverage demonstration data.

4.1. BIQN: Bayesian implicit quantile networks

In this section, we propose BIQN, a model that integrates the structures of IQN and BNN to simultaneously capture dual-mode uncertainty.

The key to constructing BIQN lies in addressing two fundamental challenges: the fusion at the representation level and the fusion at the optimization level. This requires designing a unified parameterized distribution to model the dual sources of randomness, and developing a joint loss function to align the training objectives of both components. However, given the existing forms of representation and loss functions in IQN and BNN, integrating them directly is nontrivial. To overcome this, we derive the distributional representation and corresponding loss structure of BNN in the context of DRL, enabling BNN to be naturally integrated into the IQN framework and thereby achieving unified modeling and joint training.

When applying BNN to DRL, the input x_i is a state-action pair (s_i, a_i), and the output y_i is an estimate of $Q(s_i, a_i)$, and Q-value becomes a probability disrtibution:

$$Q(s,a) = f(s,a;\theta), \quad \theta \sim q(\theta) \tag{6}$$

When training by minimizing the squared error, it is typically assumed that the error around the target value follows a Gaussian distribution. Under this assumption, the posterior distribution over the parameters in Eq. (5) can be written as:

$$\log p\big(y_i \mid x_i; \theta\big) = \frac{-\big(\delta_t(\theta_i)\big)^2}{2\sigma^2} + C(\sigma) \tag{7}$$

where $C(\sigma) = -\log\sqrt{(2\pi)}\sigma$ is constant, $\delta_t(\theta)$ is TD error at step t:

$$\delta_t(\theta) = r_t + \gamma f\big(s_{t+1}, \pi^*(s_{t+1}); \theta^-\big) - f(s_t, a_t; \theta), \quad \theta \sim q(\theta) \tag{8}$$

the sample-based $Q^\pi(s_t, a_t)$ estimation is calculated by drawing K_θ samples from $\theta \sim q(\theta)$:

$$\tilde{Q}^\pi(s,a) = \frac{1}{K_\theta}\sum_{k'=1}^{K_\theta} f(s,a;\theta_{k'}) \tag{9}$$

We approximate the integral for each example with a Monte Carlo estimate by sampling a $\hat{\theta}_i \sim q(\theta)$:

$$ELBO \approx C_1\left(\sum_{i=1}^{M} -\big(\delta_t(\hat{\theta}_i)\big)^2\right) - KL(q(\theta)|p(\theta)) \tag{10}$$

where $C_1 = 2\sigma^2$ is constant. Since DRL agents are typically trained over millions of interactions, we assume that the log-likelihood term dominates in the ELBO. Therefore, maximizing Eq. (10) is equivalently written as minimizing the following loss function:

$$\mathcal{L}_{BNN}(\hat{\theta}_i) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\frac{1}{M}\sum_i \big(\delta_t(\hat{\theta}_i)\big)^2\right] \tag{11}$$

each sample $\hat{\theta}_i \sim \mathcal{N}(\mu, \Sigma)$ is obtained by reparameterizing the network parameters: $\hat{\theta}_i = \mu + \Sigma\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$.

Building on this, we can clearly derive the overall learning procedure of BIQN. Specifically, combine Eqs. (1) and (6), BIQN uses the conditional quantile function $f_\tau(s,a;\theta)$ to represent the return variable under quantile point $\tau \sim \mathcal{U}(0,1)$ and network parameters $\theta \sim q(\theta)$:

$$Z_{\theta,\tau}(s,a) = f_\tau(s,a;\theta), \quad \tau \sim \mathcal{U}(0,1), \theta \sim q(\theta) \tag{12}$$

Combine Eqs. (2) and (8), the TD error is used to update the target value, defined as:

$$\delta_t^{\tau,\tau'}(\theta) = r_t + \gamma Z_{\theta,\tau'}\big(s_{t+1}, \pi^*\big(s_{t+1}\big)\big) - Z_{\theta,\tau}\big(s_t, a_t\big) \tag{13}$$

where $\pi^*(s_t) = \arg\max_a Q(s_t, a_t)$ is the optimal policy.

Combine Eqs. (3) and (9), the sample-based $Q^\pi(s_t, a_t)$ estimation is calculated by drawing $K_\tau \times K_\theta$ samples from $\tau \sim \mathcal{U}(0,1), \theta \sim q(\theta)$ as follows:

$$\tilde{Q}^\pi(s,a) = \frac{1}{K_\theta}\frac{1}{K_\tau}\sum_{k'=1}^{K_\theta}\sum_{j=1}^{K_\tau} Z_{\theta_{k'},\tau_k}(s,a) \tag{14}$$

Combine Eqs. (4) and (11), the training objective of BIQN is to minimize the Huber quantile loss under double sampling, defined as:

$$\mathcal{L}_{BIQN}(\theta) = \mathbb{E}_{s_t \sim \mathcal{D}}\left[\frac{1}{M}\frac{1}{N}\sum_m \sum_{i,j} \rho_\kappa\left(\delta_t^{\tau_i,\tau'_j}(\hat{\theta}_m)\right)\right] \tag{15}$$

where $\mathcal{D}$ is the replay buffer that stores training data, each $\hat{\theta}_m$ is sampled by $q(\theta) = \mathcal{N}(\mu, \Sigma)$.

4.2. Risk estimation: Metric of dual-mode uncertainty

In this section, we present the quantification and unbiased estimation of dual-mode uncertainty in BIQN.

We define the dual-mode uncertainty of a state-action pair (s,a) as the variance of the stochastic return variable $Z_{\theta,\tau}(s,a)$, which is generated by the proposed BIQN. In BIQN, the return variable is modeled as a function of both parameter uncertainty $\theta \sim q(\theta)$ and quantile variability $\tau \sim \mathcal{U}(0,1)$. Formally, the dual-mode uncertainty is defined as:

$$DU(s,a) := \mathrm{Var}[Z_{\theta,\tau}] = \mathbb{E}[Z_{\theta,\tau}^T Z_{\theta,\tau}] - \mathbb{E}[Z_{\theta,\tau}]^T \mathbb{E}[Z_{\theta,\tau}] \tag{16}$$

We model the stochastic return distribution $p(Z_{\theta,\tau}|s,a)$ as:

$$p(Z_{\theta,\tau}|s,a) = \int p(Z_{\theta,\tau}, \theta, \tau|s,a)q(\theta)p(\tau)d\theta d\tau \tag{17}$$

which:

$$\begin{cases} p(Z_{\theta,\tau}, \theta, \tau|s,a) = \mathcal{N}(f_\tau(s,a;\theta), \sigma_z). \\ q(\theta) = \mathcal{N}(\mu, \sum) \\ p(\tau) = \mathcal{U}(0,1) \end{cases} \tag{18}$$

where σ_z reflects the intensity of noise perturbation on the neural network output $f_\tau(s,a;\theta)$.

The expected return is computed by marginalizing over θ and τ:

$$\begin{aligned} E[Z_{\theta,\tau}] &= \int Z_{\theta,\tau}p(Z_{\theta,\tau}|s,a)dZ_{\theta,\tau} \\ &= \int Z_{\theta,\tau}\int\int p(Z_{\theta,\tau},\theta,\tau|s,a)q(\theta)p(\tau)d\theta d\tau \\ &= \int\int\int Z_{\theta,\tau}\mathcal{N}(f_\tau(s,a;\theta),\sigma_z)dZ_{\theta,\tau}q(\theta)p(\tau)d\theta d\tau \\ &= \int\int f_\tau(s,a;\theta)q(\theta)p(\tau)d\theta d\tau \\ &\approx \frac{1}{M}\frac{1}{M'}\sum_{i=1}^{M}\sum_{j=1}^{M'} f_{\tau_j}(s,a;\theta_i) \end{aligned} \tag{19}$$

To quantify the spread of the return distribution, we compute the second-order moment:

$$\begin{aligned} &E[Z_{\theta,\tau}^T Z_{\theta,\tau}] \\ &= \iint \Big(\int Z_{\theta,\tau}^T Z_{\theta,\tau}p(Z_{\theta,\tau},\theta,\tau|s,a)dZ_{\theta,\tau}\Big)q(\theta)p(\tau)d\theta d\tau \\ &= \int\int (cov[Z_{\theta,\tau}] + E[Z_{\theta,\tau}]^T E[Z_{\theta,\tau}])q(\theta)p(\tau)d\theta d\tau \\ &= \int (\sigma_z + f_\tau^T(s,a,\theta)f_\tau(s,a;\theta))q(\theta)p(\tau)d\theta d\tau. \\ &\approx \sigma_z + \sum_{i=1}^{M}\sum_{j=1}^{M'} f_{\tau_j}^T(s,a;\theta_i)f_{\tau_j}(s,a;\theta_i) \end{aligned} \tag{20}$$

5. HiRIL: Human-in-the-loop risk-aware imitation-enhanced learning

In this section, we propose a HiRIL method that integrates human guidance into risk-aware policy optimization. The method enhances the safety and efficiency of policy learning through two key modules: (1) a human intervention mechanism based on dual-mode uncertainty, which identifies high-risk states to trigger human takeover and uses reward shaping to guide the agent away from risky behaviors; and (2) a risk-aware prioritized experience replay mechanism incorporating human demonstrations, which introduces the Temporal Difference Risk Difference (TDRD) to improve the utilization efficiency of critical experiences and jointly optimizes the policy with imitation learning objectives.

5.1. Risk-aware human intervention

In this section, we present in detail the human-intervention-based Actor–Critic RL interaction mechanism and the associated reward shaping techniques.

In standard RL, during interaction with the environment, the agent's behavior policy $\pi_\phi(s_t)$ outputs actions to explore the environment. For Actor–Critic methods, this process can be expressed as

$$a_t^{\mathrm{RL}} = \pi_\phi(s_t) + \xi_a \odot a_t^{\mathrm{std}}, \tag{21}$$

where $a_t^{\mathrm{std}} \in \mathbb{R}^{\dim(\mathcal{A})}$ is a training-related variable that scales the exploration noise, $\odot$ denotes the Hadamard (element-wise) product, and $\xi_a \sim \mathcal{N}\big(0, I^{\dim(\mathcal{A})}\big)$.

In order to ensure that human intervention is requested only when necessary, we design a risk-aware intervention mechanism based on dual-mode uncertainty estimation. Specifically, when the agent's dual-mode uncertainty estimate $\mathrm{DU}(s_t, a_t)$ for the current state-action pair (s_t, a_t) exceeds a preset threshold τ_{risk}, the current state is regarded as lying in a high-risk region and the system triggers human intervention; at this moment, full control authority is handed over to the human. The executed action is

$$a_t = \big(I^{\dim(\mathcal{A})} - \Delta_t\big) \cdot a_t^{\mathrm{RL}} + \Delta_t \cdot a_t^H, \tag{22}$$

where a_t^H denotes the human control action, and $\Delta_t \in \mathbb{R}^{\dim(\mathcal{A})}$ is the intervention function defined by the risk state:

$$\Delta_t = \begin{cases} I^{\dim(\mathcal{A})}, & \text{if } \mathrm{DU}(s_t, a_t^{\mathrm{RL}}) > \tau_{\mathrm{risk}}, \\ \mathbf{0}^{\dim(\mathcal{A})}, & \text{otherwise}. \end{cases} \tag{23}$$

Each interaction yields a transition tuple ζ that, after the action is dispatched to the environment, is recorded and stored in the experience replay buffer $\mathcal{D}$. In particular, the actions produced by both the human policy and the RL policy are associated with a dual-mode uncertainty estimate DU_t and the intervention function Δ_t. Accordingly, the new transition tuple ζ_i is defined as follows to distinguish human experience from ordinary RL experience:

$$\zeta_i = (s_i, a_i, r_i, s_{i+1}, DU_i, \Delta_i). \tag{24}$$

Clearly, when an intervention occurs, the current state is high risk for the RL agent; in this context, the intervention event can be regarded as a negative signal from which the agent should learn to avoid the state. In reinforcement learning, the agent updates its value function and policy based on the reward obtained at each interaction; therefore, the influence of human intervention can be encoded via reward shaping. Specifically, the post-intervention reward is defined as

$$r_t^{\mathrm{shape}} = r_t + r_{\mathrm{p}} \cdot \big[(\Delta_t = \mathbf{I}^{\dim(\mathcal{A})}) \wedge (\Delta_{t-1} = \mathbf{0}^{\dim(\mathcal{A})})\big], \tag{25}$$

where r_{p} is a coefficient that weights the intervention penalty. Note that the shaping penalty is applied only at the onset of intervention, i.e., $(\Delta_t = \mathbf{I}^{\dim(\mathcal{A})}) \wedge (\Delta_{t-1} = \mathbf{0}^{\dim(\mathcal{A})})$, because subsequent states are governed by human actions and should no longer be treated as high-risk for the RL agent.

5.2. Risk-aware prioritized experience replay and imitation learning

In this section, we propose a risk-aware prioritized experience replay mechanism and imitation learning objectives tailored to human demonstrations.

In RL, data are typically sampled from the replay buffer uniformly; however, this treats every sample equally and cannot fully exploit the value of different experiences. A more effective method is Prioritized Experience Replay (PER), which assumes that experiences in the buffer $\mathcal{D}$ follow a certain distribution $\mathcal{J}$, whose probability mass function is defined as:

$$p_J(i) = \frac{k_i}{\sum_{i \in \mathcal{D}} k_i}. \tag{26}$$

The priority is determined by the temporal-difference (TD) error δ_i^{TD}, which is computed as follows:

$$\begin{aligned} k_i &= \left|\delta_i^{TD}\right| + \varepsilon \\ &= \left| r_i + \gamma \cdot Q\big(s_{i+1}, \pi_\phi(s_{i+1}); \theta\big) - Q\big(s_i, a_i; \theta\big)\right| + \varepsilon \end{aligned} \tag{27}$$

where ε is a small positive constant to guarantee the probability larger than zero.This formulation indicates that higher state risk yields a lower base priority, thereby reducing the sampling frequency of high-risk data.

When the experience comes from human demonstrations, we augment the sampling priority defined in Eq. (29) with an exponential term representing the difference between the DU values of the human action and the RL action. This additional component is referred to as the Risk Difference (RD) term, which indicates that when the human action significantly reduces risk, the priority of the corresponding sample is increased. The improved priority is then defined as

$$k_i^H = \left|\delta_i^{TD}\right| + \varepsilon + (\Delta_i = \mathbf{I}^{\dim(\mathcal{A})}) \cdot \exp\big[\mathrm{DU}(s_i, a_i^{\mathrm{RL}}) - \mathrm{DU}(s_i, a_i^H)\big]. \tag{28}$$

We refer to the aforementioned mechanism as **TDRD**.

Assume that the replay buffer $\mathcal{D}$ is divided into RL experiences and human experiences, denoted as $\mathcal{D}_1 \cup \mathcal{D}_2$. The critic loss is defined as

$$\begin{aligned} \mathcal{L}_{\mathrm{critic}}(\theta) = \mathbb{E}_{\mathcal{D}_1}&\left[\frac{1}{M}\frac{1}{N}\sum_m \sum_{i,j} \rho_\kappa\Big(\delta_t^{\tau_i,\tau_j',1}(\hat{\theta}_m)\Big)\right] \\ &+ \mathbb{E}_{\mathcal{D}_2}\left[\frac{1}{M}\frac{1}{N}\sum_m \sum_{i,j} \rho_\kappa\Big(\delta_t^{\tau_i,\tau_j',2}(\hat{\theta}_m)\Big)\right], \end{aligned} \tag{29}$$

where

$$\delta_t^{\tau,\tau',1}(\theta) = r_t + \gamma Z_{\theta,\tau'}\big(s_{t+1}, \pi^*(s_{t+1})\big) - Z_{\theta,\tau}\big(s_t, a_t^{RL}\big), \tag{30}$$

$$\delta_t^{\tau,\tau',2}(\theta) = r_t + \gamma Z_{\theta,\tau'}\big(s_{t+1}, \pi^*(s_{t+1})\big) - Z_{\theta,\tau}\big(s_t, a_t^{H}\big). \tag{31}$$

The actor loss is defined as:

$$\mathcal{L}_{\mathrm{Actor}}(\phi) = \mathbb{E}_{\mathcal{D}_1}\big[-\tilde{Q}(s_t, \pi_\phi(s_t))\big] + \lambda \cdot \mathbb{E}_{\mathcal{D}_2}\big[\|\pi_\phi(s_t) - a_t^H\|^2\big]. \tag{32}$$

where λ is a manually determined constant that weighs the importance of behavior cloning.

In summary, the complete form of the algorithm is presented in Algorithm 1.

6. Simulation

We conduct experiments to investigate the following questions: (1) Whether HiRIL can further improve the efficiency and performance of reinforcement learning during the training phase compared to baseline methods; (2) Whether agents trained with HiRIL exhibit enhanced robustness and adaptability during the testing phase compared to baselines; (3) Whether the mechanism design of HiRIL is rational and effective. For Question (1), we comprehensively compare the training efficiency and performance of different algorithms under the same hyperparameter settings using multiple evaluation metrics. For Question (2),

Algorithm 1 HiRIL algorithm.

```
**Initialize:** Replay buffer $\mathcal{D}$; Parameters $\theta = (\mu, \Sigma)$ and $\phi$ randomly;
  Threshold $\tau_{\text{risk}}$ for intervention triggering.
**for** each episode **do**
    $s_0 \leftarrow$ initial state
    **for** $t = 0$ to $T$ **do**
        $a_t^{\text{RL}} \leftarrow \pi_\phi(s_t) + \epsilon$;
        Estimate $\text{DU}(s_t, a_t^{\text{RL}})$ using Eq. (16);
        **if** $\text{DU}(s_t, a_t^{\text{RL}}) > \tau_{\text{risk}}$ **then**
            Adopt human action $a_t = a_t^{\text{H}}$, set $\Delta_t = I$;
        **else**
            Select RL action $a_t = a_t^{\text{RL}}$, set $\Delta_t = 0$;
        Execute $a_t$, observe $r_t$ and new state $s_{t+1}$;
        Shape reward $r_t = r_t + r_{\text{p}} \cdot \left[(\Delta_t = \mathbf{I}) \wedge (\Delta_{t-1} = \mathbf{0})\right]$;
        Store tuple $(s_t, a_t, r_t, s_{t+1}, DU_t, \Delta_t)$ in $\mathcal{D}$;
        Sample $N$ tuples from $\mathcal{D}$ with probability $p(i) = \frac{k_i}{\sum_{i \in \mathcal{D}} k_i}$;
        Update priority by Eq. (28);
        /* Critic Update */
        Compute critic loss $\mathcal{L}_{\text{Critic}}(\theta)$ using Eq. (29);
        Update $\theta$ via gradient descent;
        /* Actor Update */
        Estimate $\hat{Q}(s, a)$ using Eq. (14);
        Compute actor loss $\mathcal{L}_{\text{Actor}}(\phi)$ using Eq. (32);
        Update $\phi$ via gradient descent;
```

we evaluate the agents from three perspectives: adaptability to different environments, robustness to non-stationary human guidance, and robustness to control input noise. For Question (3), we mainly verify whether the HiRIL mechanism can effectively trigger interventions and reduce the frequency of human interventions, and we conduct a contribution analysis of each module.

6.1. Environment setup

As with most RL algorithms, the proposed HiRIL can be broadly applied to decision-making and control tasks with continuous action spaces. In this paper, we focus on end-to-end autonomous driving as the research domain and adopt the CARLA simulator as the experimental platform, since CARLA can generate an unlimited number of scenarios across diverse road networks and traffic flows.

We set up six representative scenarios: one for training the proposed method and the remaining five for testing and evaluating its performance. Visualizations of all scenarios are provided in Fig. 2. In the training scenario, seven surrounding vehicles (all sedans) are placed around the ego vehicle. Their initial speeds range between [3, 5] m/s, and subsequent acceleration is governed by the IDM model (Treiber et al., 2000). The initial positions of all vehicles are fixed at the beginning of each training episode. The differences between training and testing scenarios lie in the number of surrounding vehicles, their initial position distributions, and vehicle types. In these scenarios, the task objective is to drive the autonomous vehicle safely to its destination while avoiding hazardous behaviors such as collisions with other vehicles or lane departures. Reward shaping includes dense rewards proportional to lateral control stability, a terminal reward when the ego vehicle successfully reaches the destination, and penalties for collisions or lane departures.

We adopt state-of-the-art algorithms in the field of HiL-RL as baselines and compare their performance with our proposed algorithm: **IARL** (Wang et al., 2018): This is a representative method that combines reinforcement learning with imitation learning. Specifically, the RL policy network is modified by incorporating a behavior cloning objective to adapt to human demonstration actions. Once human intervention occurs, human demonstrations replace RL exploratory actions, and a penalty signal is added to the reward. **HULA** (Singi et al., 2024): This is a representative method that combines reinforcement learning with human intervention. It requests assistance from human experts by associating decision uncertainty with the return variance of the agent's perceived current state, without modifying the network structure or optimization procedure. **DRL**: In this experiment, we employ the TD3 (Fujimoto et al., 2018) algorithm as the baseline RL model. Additionally, we implement the PER mechanism in all the above baselines to ensure a fair comparison. The specific hyperparameters used during training are listed in Tables A1 and A2 in the Appendix. For baseline algorithms involving human participation, we assume that the human operator possesses professional operational proficiency. Before the experiment begins, each participant is required to independently control the vehicle for 20 episodes in the training scenario to become familiar with the operation process. Human operators are involved throughout the training process, while in the testing scenarios, only the trained policies are evaluated without human involvement.

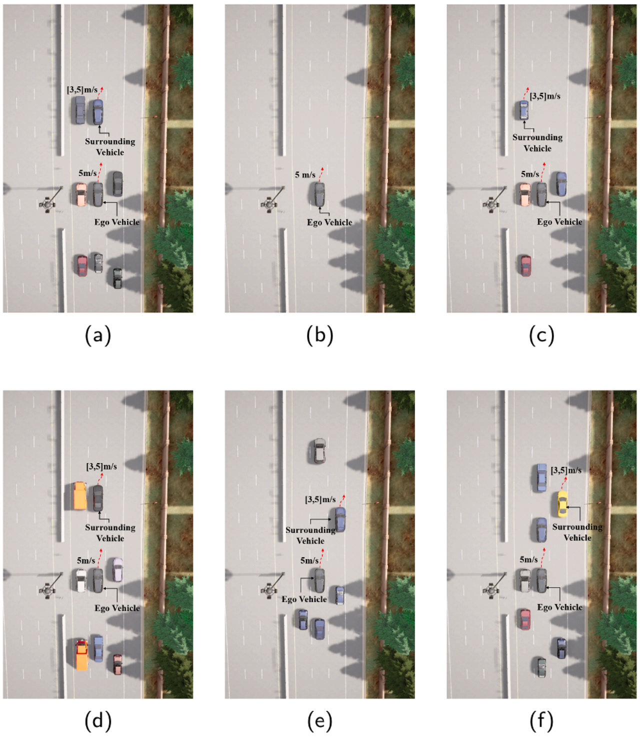

Fig. 2. (a) Scenario 0: This scenario serves as the training scenario. (b) Scenario 1: This scenario removes all surrounding traffic participants to evaluate the anti-overfitting capability of the generated driving policy. (c–f) Scenarios 2–5: These scenarios are used to test the adaptability of the obtained policy in new situations not encountered during the training phase. Changes include the number of surrounding vehicles, initial positions, and vehicle types.

We employ four metrics to evaluate learning performance: **Reward**: the cumulative reward obtained by the agent in each episode (excluding human intervention shaping rewards). **Surviving distance**: the distance traveled by the ego vehicle before reaching either the goal state or a failure state. **Success rate**: the proportion of episodes in which the agent successfully completes the task in the testing environments. **Collision rate**: the proportion of episodes in which the ego vehicle collides with obstacles or deviates from the route.

6.2. Training performance evaluation

In this section, we verify whether the proposed HiRIL method demonstrates superior training performance compared to other state-of-the-art HiL algorithms. The evaluation focuses on both learning performance and safety. Learning performance is assessed using reward and survival distance, while safety is evaluated based on the collision rate during training.

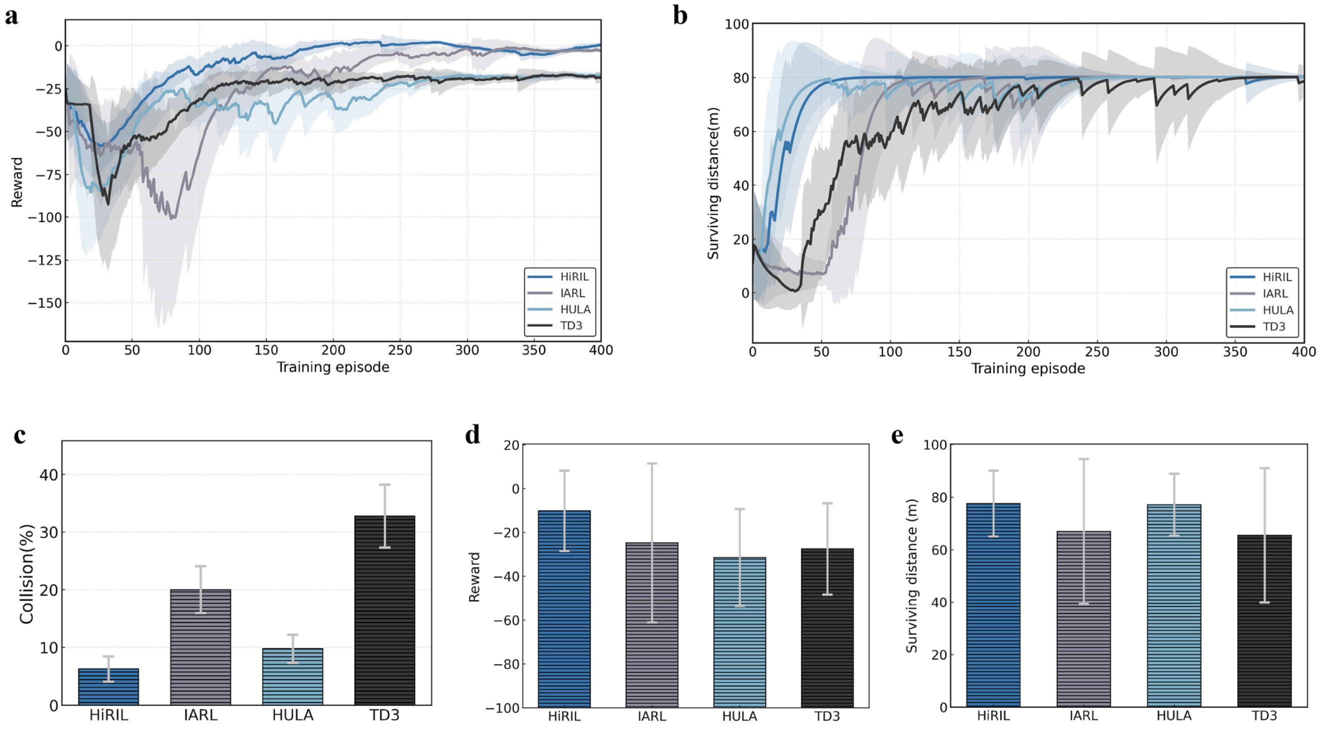

Fig. 3. Comparison of training performance across four different methods (HiRIL, IARL, HULA, and TD3): a. Episodic training rewards during the training process, with mean and standard deviation calculated based on four different random seeds; **b.** Episodic surviving distances during the training process, with mean and standard deviation calculated based on four different random seeds; **c.** Collision rates throughout the training process, with mean and standard deviation calculated based on four different random seeds; **d.** Average rewards across all episodes, with mean and standard deviation calculated over 400 episodes; **e.** Average surviving distances across all episodes, with mean and standard deviation calculated over 400 episodes.

Fig. 3(a) and (b) visualize the learning performance in the form of curves, where solid lines represent the mean values and the shaded areas indicate the standard deviation (all algorithms are trained with four different random seeds). In Fig. 3(a), HiRIL achieves faster reward convergence and higher final rewards than the other methods, indicating superior training efficiency and policy performance. Fig. 3(b) shows the evolution of survival distance during training. The results reveal that the standard TD3 algorithm struggles to improve the policy, with frequent fluctuations even in the later training stages. In contrast, the three HiL algorithms perform better, with faster learning rates—especially HiRIL, which achieves the highest survival distance among all baseline algorithms in around 50 episodes. The evaluation of computational efficiency is shown in Table A3.

Fig. 3(d) and (e) present statistical comparisons of rewards and survival distances over 400 training episodes. The bars indicate the mean values, and the error bars represent the standard deviations. From Fig. 3(c), it can be seen that HiRIL achieves the highest average reward over the entire training process (M = −10.14, SD = 18.37), followed by IARL (M = −24.79, SD = 36.28), TD3 (M = −27.53, SD = 20.88), and HULA (Mean = −31.46, SD = 32.08). As shown in Fig. 2(d), HiRIL also achieves the longest average survival distance (M = 77.57, SD = 10.8), followed by HULA (M = 77.15, SD = 12.5), IARL (M = 66.94, SD = 31.56), and TD3 (M = 65.48, SD = 25.53).

To assess safety, we calculate the collision rates of the ego vehicle over 400 episodes, as shown in Fig. 3(c). HiRIL reports the lowest collision rate (M = 6.25%, SD = 2.18), significantly outperforming IARL (M = 20.00%, SD = 4.06), HULA (M = 9.75%, SD = 2.42), and TD3 (M = 32.75%, SD = 5.45). These results indicate that HiRIL not only improves policy performance but also enhances safety during the training process.

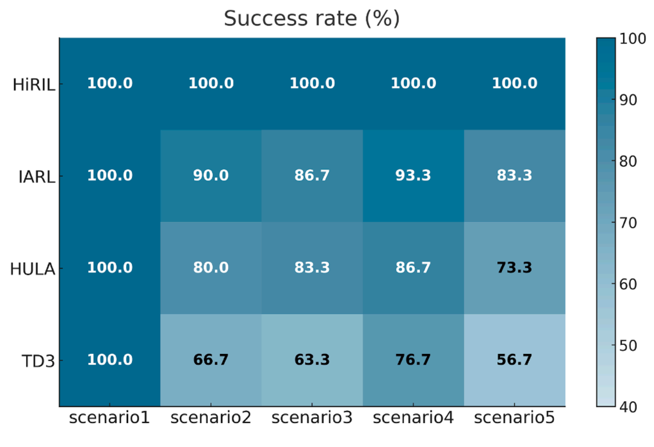

Fig. 4. Success rates of policies trained with different methods across the five testing scenarios.

6.3. *Testing performance evaluation*

In this section, we evaluate the practicality of the aforementioned algorithms by testing the trained policies in terms of safety and robustness. For safety, we use the success rate as the evaluation metric to assess whether the policies can effectively avoid high-risk behaviors in scenarios not encountered during training. For robustness, we conduct evaluations from the following three perspectives: **Adaptability to new environments**: This assesses whether the policy can adapt well to scenarios different from those used during training. Specifically, we modify the number of vehicles, their initial positions, and vehicle types in the test environment to create clear discrepancies from the training conditions. **Robustness to non-stationary human guidance**: This evaluates the policy's stability under demonstrations of varying quality. We simulate two types of non-stationary human operators: The first type represents operators affected by unexpected factors, resulting in occasional erroneous operations. This is simulated by replacing 1/4 of the human demonstration data in the buffer with random actions. The second type represents inexperienced beginners, simulated by providing no prior training before the experiment begins. **Robustness to noisy disturbances**: To simulate external interference such as sensor errors or actuator jitter, we inject Gaussian noise with zero mean and a standard deviation equal to 5% of the control range into the control commands. This evaluates the policy's sensitivity to low-level input disturbances.

To evaluate safety, we repeated each experiment 30 times in every testing scenario using the same sequence of random seeds. As shown in Fig. 4, the agent trained with HiRIL successfully completed all previously unseen tasks, whereas the baseline methods succeeded only in part of them. The reason HiRIL achieves the highest success rate among all methods is that it leverages human interventions more efficiently at both the value-learning and policy-learning levels. Specifically, HiRIL triggers human guidance based on dual uncertainty, which focuses demonstration data on safety-critical states with high uncertainty. By combining behavior cloning loss with a risk-aware PER mechanism, HiRIL repeatedly updates the policy on these human-corrected experiences, enabling it to utilize human interventions more precisely and guide the policy toward safer and more generalizable behaviors.

To evaluate the robustness of different methods in handling variations in the quality of human guidance, we use the average survival distance as the performance metric. A smaller performance drop under non-stationary human guidance is considered indicative of higher robustness. As shown in Fig. 5, HiRIL exhibits the smallest performance fluctuations across all testing scenarios, with its results under both "stationary" and "non-stationary" conditions mostly close to the balanced 50:50 dividing line. This indicates that the learned policy is more robust to changes in human guidance quality. Since IARL adopts a behavior cloning strategy to imitate human guidance, it is more susceptible to the negative impact of low-quality demonstrations. In contrast, HULA relies solely on human interventions without imitation, making it less affected by the quality of guidance. Moreover, the results also show that all three methods demonstrate better robustness when dealing with occasionally erroneous human operators compared to inexperienced ones. This suggests that a basic level of operational competence in human demonstrations has a greater influence on policy training than occasional mistakes.

We also evaluated the robustness of each method under noisy conditions. Specifically, disturbances were injected into the control commands, and the performance was assessed across five different scenarios. As shown in Table 1, HiRIL achieved the highest task distance (79.10 ± 1.54) and success rate (93.3 ± 2.4) under perturbations, outperforming IARL, HULA, and TD3. IARL performed slightly better than HULA, primarily because it incorporates a behavior cloning objective during human intervention, which allows the policy to align more closely with expert behavior in critical states. As a result, it learns a more conservative and noise-robust policy.

Table 1
Comparison of survival distance and success rate for four methods in noise-injected scenarios.

Method	Survival Distance	Success Rate
HiRIL	**79.10±(1.54)**	**93.3±(2.4)**
IARL	77.82±(1.90)	81.3±(3.8)
HULA	75.54±(3.29)	78.0±(3.8)
TD3	73.18±(5.02)	58.7±(5.1)

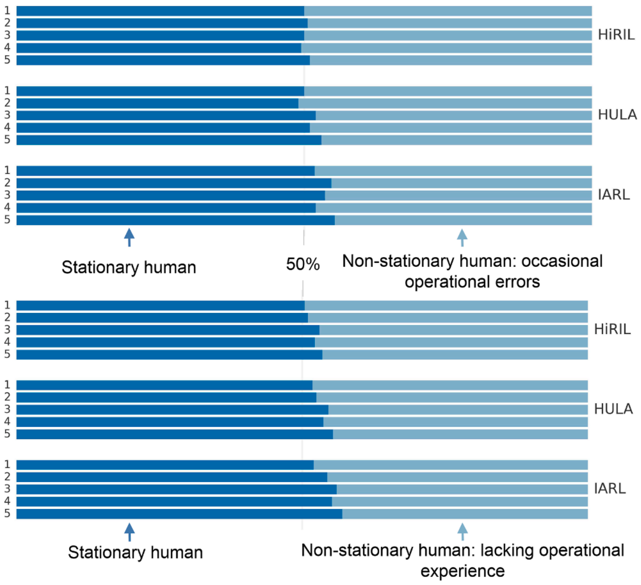

Fig. 5. Bar chart of survival distances in five test environments under the guidance of stationary humans and non-stationary humans (including two types: those with occasional operational errors and those lacking operational experience).

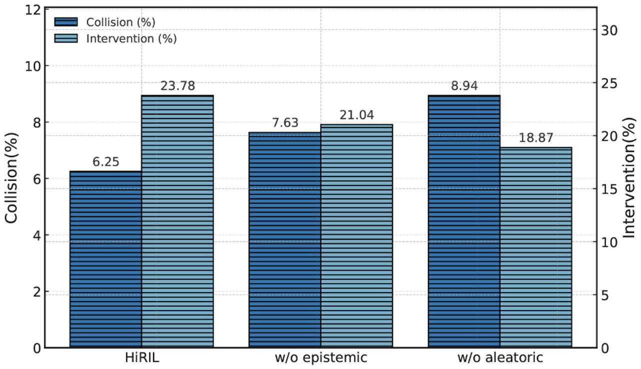

Fig. 6. Comparison of collision rates and intervention rates with or without epistemic or aleatoric uncertainty.

6.4. Validation of mechanism effectiveness

This section evaluates the effectiveness of the proposed method from three perspectives: (i) whether the HiRIL can effectively trigger interventions; (ii) whether HiRIL can effectively reduce the frequency of human interventions; (iii) the performance contribution of each key module.

Effectiveness Analysis of HiRIL in Triggering Interventions. Within the HiRIL framework, we conduct comparative ablation experiments by removing either the aleatoric uncertainty described in Section 3.1 or the epistemic uncertainty described in Section 3.2. As shown in Fig. 6, removing either source of uncertainty results in higher collision rates and lower intervention rates, indicating a weakened risk estimation capability. Therefore, the dual-mode uncertainty modeling enables a more precise intervention-triggering mechanism and achieves a better balance between safety and efficiency.

Effectiveness Analysis of HiRIL in Reducing Intervention Frequency. We first verify the relationship between the collision rate and intervention rate under different risk thresholds, and then use step-based and episode-based intervention rates to examine whether the frequency of human interventions decreases as training progresses. As shown in Fig. 7(a), as the risk threshold decreases, interventions are triggered more frequently. Specifically, the intervention rate

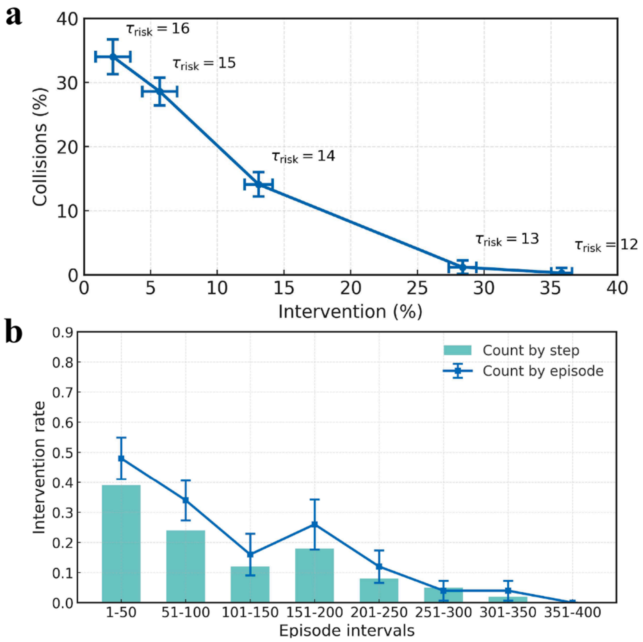

Fig. 7. **(a)** Relationship between collision rate (%) and intervention rate (%) under different risk thresholds. with the mean and standard deviation computed over four different random seeds. **(b)** Human intervention rates throughout the entire training process of HiRIL. Two metrics are used for evaluation: "step-based" and "episode-based" intervention rates. The step-based metric calculates the proportion of time steps guided by humans within a given episode interval, with its standard deviation computed over 50 episodes in that interval; the episode-based metric denotes the proportion of episodes within the interval in which human intervention occurs.

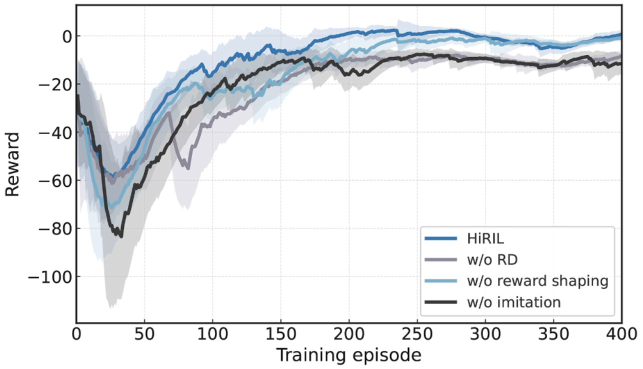

Fig. 8. Ablation study of HiRIL. Training reward curves under different configurations: full HiRIL, without risk difference (w/o RD), without reward shaping, and without imitation. The mean and standard deviation are calculated based on four different random seeds.

increases from $2.18\% \rightarrow 5.68\% \rightarrow 13.10\% \rightarrow 28.38\% \rightarrow 35.81\%$, while the collision rate drops significantly from $34.0\% \rightarrow 28.6\% \rightarrow 14.1\% \rightarrow 1.2\% \rightarrow 0.3\%$. This trend indicates that lower risk thresholds help trigger interventions more promptly in high-risk situations, thereby effectively reducing collisions. However, more frequent interventions may increase human workload and operational costs. Therefore, in practical applications, it is important to strike a balance between safety and intervention cost by selecting an appropriate risk threshold. Fig. 7(b) shows the change in human intervention rates during the training process of HiRIL. The results reveal a clear downward trend in intervention frequency, suggesting that the agent gradually learns to behave more safely and robustly, becoming increasingly less dependent on human guidance. This trend demonstrates the effectiveness of the imitation mechanism in improving training efficiency and reducing the cost of human intervention.

Module Contribution Analysis. We conducted ablation experiments to analyze the contribution of different modules to performance improvement, and the results are shown in Fig. 8. In Fig. 8, the three ablated modules correspond to: the intervention-based reward shaping mechanism in Eq. (28), the risk-difference (RD) based prioritized experience replay in Eq. (31), and the behavior cloning objective in Eq. (35). The experimental results indicate that the imitation learning objective and reward shaping mechanism play a critical role in enhancing algorithm performance. Removing either of these modules leads to a significant reduction in training efficiency and final performance. In contrast, removing the risk-difference (RD) term has a relatively minor impact.

7. Conclusion

This paper proposes HiRIL, a human-in-the-loop reinforcement learning method that integrates a risk-aware intervention-triggering mechanism with imitation-based policy optimization, aiming to improve sample efficiency and policy safety in safety-critical continuous control tasks. The approach models epistemic and aleatoric uncertainties using BIQN and constructs a quantitative risk metric by providing an unbiased estimation of the second-order moment of their joint distribution. This risk measure enables effective human intervention triggering and efficient utilization of human demonstration data. Experimental results show that HiRIL significantly outperforms baselines on the CARLA end-to-end autonomous driving benchmark and exhibits excellent robustness and generalization capabilities under challenging conditions such as input noise and non-stationary human interventions.

However, the proposed method still has several limitations, mainly in the following three aspects: The intervention triggering conditions do not explicitly account for suboptimal human policy performance, which may limit the policy function due to human performance ceilings; The intervention threshold is set as a fixed value, this static rule may become ineffective when facing distributional shifts or changes in task complexity; A fixed weight is used when incorporating human intervention data into policy updates, neglecting both the variability among human participants and the agent's continuously improving capabilities. Future work could explore these three directions to further enhance the adaptability and practicality of the proposed method.

CRediT authorship contribution statement

Yaqing Zhou: Conceptualization, Methodology, Data curation, Writing – Original draft preparation; **Yun-Bo Zhao:** Conceptualization, Supervision, Funding acquisition; **Chenwei Xu:** Visualization, Software; **Chen Ouyang:** Investigation; **Pengfei Li:** Validation, Writing – Review & Editing.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62473352).

Appendix A.

(Table A1-A3).

Table A1
Default values of training hyperparameters used in the experiments.

Hyperparameter	Value
Batch Size	128
Max Training Epochs	400
Replay buffer size	38,400
Initial exploration	1
Final exploration	0.05
Discount Factor	0.95
Value Network Learning Rate	0.0005
Policy Network Learning Rate	0.0002
Policy Update Delay Frequency	2
k_τ	8
k_θ	8
Huber Loss Threshold κ	1.0
Number of Convolutional Layer Channels	(6, 16)
Fully Connected Layer Parameters	(256, 128, 64, 2)

Table A2
Hyperparameters for the PER mechanism.

Type	Value
Priority factor	1
Sample factor	1
Offset factor (ϵ)	10^{-3}

Table A3
Comparation of computational cost per 200 steps.

Algorithm	Time consumption (s) per 200 steps
HiRIL	6.50
HULA	6.18
IARL	6.32
TD3	6.08

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

Celemin, C., Pérez-Dattari, R., Chisari, E., Franzese, G., de Souza Rosa, L., Prakash, R., Ajanović, Z., Ferraz, M., Valada, A., Kober, J. et al. (2022). Interactive imitation learning in robotics: A survey. *Foundations and Trends in Robotics*, *10*(1-2), 1–197.

Clements, W. R., Delft, B. V., Robaglia, B.-M., Slaoui, R. B., & Toth, S. (2019). Estimating risk and uncertainty in deep reinforcement learning. arXiv preprint arXiv:1905.09638.

Dabney, W., Ostrovski, G., Silver, D., & Munos, R. (2018a). Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th international conference on machine learning (ICML)* (pp. 1096–1105). PMLR.

Dabney, W., Rowland, M., Bellemare, M., & Munos, R. (2018b). Distributional reinforcement learning with quantile regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 2892–2901.

Fujimoto, S., & Gu, S. S. (2021). A minimalist approach to offline reinforcement learning. *Advances in Neural Information Processing Systems*, *34*, 20132–20145.

Fujimoto, S., Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. In *International conference on machine learning* (pp. 1587–1596). PMLR.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., & Roscher, R., et al. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, *56*(Suppl 1), 1513–1589.

Goan, E., & Fookes, C. (2020). Bayesian neural networks: An introduction and survey. In *Case studies in applied bayesian data science: CIRM jean-morlet chair, fall 2018* (pp. 45–87). Springer.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I. et al. (2018). Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*. (*vol. 32*).

Hiraoka, T., Imagawa, T., Hashimoto, T., Onishi, T., & Tsuruoka, Y. (2021). Dropout q-functions for doubly efficient reinforcement learning. arXiv preprint arXiv:2110.02034.

Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, *29*, 4565–4573.

Hoel, C.-J., Wolff, K., & Laine, L. (2023). Ensemble quantile networks: Uncertainty-aware reinforcement learning with applications in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, *24*(6), 6030–6041.

Hua, J., Zeng, L., Li, G., & Ju, Z. (2021). Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning. *Sensors*, *21*(4), 1278.

Huang, W., Liu, H., Huang, Z., & Lv, C. (2024). Safety-aware human-in-the-loop reinforcement learning with shared control for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, *25*(11), 16181–16192.

Lanzaro, G., & Sayed, T. (2024). Evaluating driver-pedestrian interaction behavior in different environments via Markov-game-based inverse reinforcement learning. *Expert Systems with Applications*, *223*, 120913.

Liu, H., Nasiriany, S., Zhang, L., Bao, Z., & Zhu, Y. (2025). Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. *The International Journal of Robotics Research*, *44*(10–11), 1727–1742.

Liu, Q. (2025). CoRLHF: Reinforcement learning from human feedback with cooperative policy-reward optimization for LLMs. *Expert Systems with Applications*, *233*(0), 122385.

Lockwood, O., & Si, M. (2022). A review of uncertainty for deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence and interactive digital entertainment (AIIDE)* (pp. 155–162). (*vol. 18*).

Mandel, T., Liu, Y.-E., Brunskill, E., & Popović, Z. (2017). Where to add actions in human-in-the-loop reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*. (*vol. 31*).

Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W., & Abbeel, P. (2018). Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE International conference on robotics and automation (ICRA)* (pp. 6292–6299). IEEE.

Osband, I., Blundell, C., Pritzel, A., & Roy, B. V. (2016). Deep exploration via bootstrapped DQN. *Advances in Neural Information Processing Systems*, *29*, 4026–4034.

Retzlaff, C. O., Das, S., Wayllace, C., Mousavi, P., Afshari, M., Yang, T., Saranti, A., Angerschmid, A., Taylor, M. E., & Holzinger, A. (2024). Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities. *The Journal of Artificial Intelligence Research*, *79*, 359–415.

Saunders, W., Sastry, G., Stuhlmueller, A., & Evans, O. (2017). Trial without error: Towards safe reinforcement learning via human intervention. arXiv preprint arXiv:1707.05173.

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2015). Prioritized experience replay. arXiv preprint arXiv:1511.05952.

Silva, F. L. D., Hernandez-Leal, P., Kartal, B., & Taylor, M. E. (2020). Uncertainty-aware action advising for deep reinforcement learning agents. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 5792–5799). (*vol. 34*).

Singi, S., He, Z., Pan, A., Patel, S., Sigurdsson, G. A., Piramuthu, R., Song, S., & Ciocarlie, M. (2024). Decision making for human-in-the-loop robotic agents via uncertainty-aware reinforcement learning. In *2024 IEEE International conference on robotics and automation (ICRA)* (pp. 7939–7945). IEEE.

Tan, M., Tao, Y., Zheng, B., Xie, G., Feng, L., Xia, Z., & Xiong, J. (2025). Safe navigation for robotic digestive endoscopy via human intervention-based reinforcement learning. *Expert Systems with Applications*, *294*, 128841.

Treiber, M., Hennecke, A., & Helbing, D. (2000). Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, *62*(2), 1805–1824.

van der Vaart, P. R., Yorke-Smith, N., & Spaan, M. T. J. (2024). Bayesian ensembles for exploration in deep q-learning. In *Proceedings of the 23rd international conference on autonomous agents and multiagent systems (AAMAS '24)* (pp. 2528–2530). IFAAMAS.

Wang, F., Zhou, B., Chen, K., Fan, T., Zhang, X., Li, J., Tian, H., & Pan, J. (2018). Intervention aided reinforcement learning for safe and practical policy optimization in navigation. In *Proceedings of the conference on robot learning (coRL)* (pp. 410–421). PMLR.

Wu, J., Huang, Z., Hu, Z., & Lv, C. (2022a). Toward human-in-the-loop AI: Enhancing deep reinforcement learning via real-time human guidance for autonomous driving. *Engineering*, *21*(2), 75–91.

Wu, J., Huang, Z., Huang, W., & Lv, C. (2022b). Prioritized experience-based reinforcement learning with human guidance for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(1), 855–869.

Xie, A., Tajwar, F., Sharma, A., & Finn, C. (2022a). When to ask for help: Proactive interventions in autonomous reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 16918–16930.

Xie, A., Tajwar, F., Sharma, A., & Finn, C. (2022b). When to ask for help: Proactive interventions in autonomous reinforcement learning. *Advances in Neural Information Processing Systems*, *35*, 16918–16930.

Xu, Y., Liu, Z., Duan, G., Zhu, J., Bai, X., & Tan, J. (2022). Look before you leap: Safe model-based reinforcement learning with human intervention. In *Conference on robot learning* (pp. 332–341). PMLR.

Yu, T., & Chang, Q. (2022). User-guided motion planning with reinforcement learning for human-robot collaboration in smart manufacturing. *Expert Systems with Applications*, *209*, 118291.

Zare, M., Kebria, P. M., Khosravi, A., & Nahavandi, S. (2024). A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, *54*(12), 7173–7186.

Zhang, Q., Kang, Y., Zhao, Y.-B., Li, P., & You, S. (2021). Traded control of human–machine systems for sequential decision-making based on reinforcement learning. *IEEE Transactions on Artificial Intelligence*, *3*(4), 553–566.