Fuz-RL: A Fuzzy-Guided Robust Framework for Safe Reinforcement Learning under Uncertainty

Xu Wan
Zhejiang University
wanxu@zju.edu.cn

Chao Yang
Alibaba DAMO Academy
xiuxin.yc@alibaba-inc.com

Cheng Yang
Alibaba DAMO Academy
charis.yangc@alibaba-inc.com

Jie Song
Peking University
jie.song@pku.edu.cn

Mingyang Sun*
Peking University
smy@pku.edu.cn

Abstract

Safe Reinforcement Learning (RL) is crucial for achieving high performance while ensuring safety in real-world applications. However, the complex interplay of multiple uncertainty sources in real environments poses significant challenges for interpretable risk assessment and robust decision-making. To address these challenges, we propose **Fuz-RL**, a fuzzy measure-guided robust framework for safe RL. Specifically, our framework develops a novel fuzzy Bellman operator for estimating robust value functions using Choquet integrals. Theoretically, we prove that solving the Fuz-RL problem (in Constrained Markov Decision Process (CMDP) form) is equivalent to solving distributionally robust safe RL problems (in robust CMDP form), effectively reformulating the min-max optimization problem into a tractable CMDP with Choquet-integrated value functions. Empirical analyses on safe-control-gym and safety-gymnasium scenarios demonstrate that Fuz-RL effectively integrates with existing safe RL baselines in a model-free manner, significantly improving both safety and control performance under various types of uncertainties in observation, action, and dynamics. The code is available in https://github.com/waunx/FuzRL.

1 Introduction

While safe reinforcement learning (RL) has achieved remarkable success in safety-crucial decision-making tasks, deploying safe RL in real-world applications remains challenging due to multiple sources of uncertainty [5, 8, 42]. Recent methods using Lyapunov functions and reachability analysis provide theoretical safety guarantees for control tasks [4, 9, 45, 48, 44, 43], but focus primarily on idealized, deterministic settings. These approaches struggle with the complex, coupled uncertainties of real-world systems, including sensor noise, actuator delays, and environmental variations.

Existing robust approaches to safe RL face several key limitations for real-world tasks. Traditional min-max techniques [49, 29, 40] focus on worst-case scenarios, resulting in overly conservative policies and computational intractability. While distributionally robust methods attempt to model uncertainty distributions, they typically assume simplified, independent uncertainties through KL-divergence constraints [39] or Gaussian perturbations [41], and treat different perturbations with equal importance. Risk-sensitive approaches using probability measures like conditional Value-at-Risk (VaR) [38], Wang transform [34], and Entropic VaR [31] enable uncertainty quantification through

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

coherent risk functionals, require careful parameter tuning and struggle to handle multiple noise sources effectively.

However, when multiple uncertainties are correlated and converge on a single system component, the resultant performance degradation often exhibits super-additive behavior. To handle such coupled uncertainties, *fuzzy measure theory* has shown promise in various decision-making tasks, from robust motion planning [7, 53, 18] to adaptive control [27, 16], through its ability to model non-additive effects and provide clear behavioral interpretations. While these successes demonstrate the potential of fuzzy measures for uncertainty quantification, extending this approach to constrained Markov decision process (CMDP) remains challenging, particularly in balancing performance objectives with safety constraints under uncertainty.

Motivated by this, we propose a novel **Fuz**zy-guided framework for Safe **RL** (Fuz-RL) that unifies uncertainty quantification and enhances current safe RL's robustness through fuzzy theory. Specifically, our main contributions are:

(1) We introduce a novel fuzzy Bellman operator that integrates Choquet integrals of fuzzy measure into value function to achieve robust value estimation under potential perturbations.

(2) We provide robustness equivalence for our Fuz-RL framework by demonstrating that solving Fuz-RL problem (a CMDP form) is equivalent to distributionally robust safe RL problems (a robust CMDP form).

(3) By seamlessly integrating the Fuz-RL framework into three safe RL methods, we conduct several robust assessments involving observation, action, and dynamics uncertainty for safe-control-gym tasks and safety-gymnasium tasks. As expected, Fuz-RL significantly enhances the robustness of safe RL algorithms in multi-source uncertainty scenarios.

2 Related Work

Our work builds upon and connects two main research directions: robust approaches in safe RL and fuzzy-based uncertainty quantification in MDPs. We review relevant work in these areas and highlight the research gaps our work addresses.

Robust Approaches in Safe RL. Uncertainties within the CMDP framework manifest in various forms, including state shifts [23], action disturbances [41], and dynamics uncertainty [33]. To address these challenges, distributionally robust optimization has been employed, where policies are optimized against worst-case transition kernels within a Wasserstein ambiguity set [39, 28]. An alternative direction leverages risk-sensitive measures to enhance robustness under safety constraints. For instance, Conditional Value-at-Risk (CVaR) has been integrated into policy optimization to explicitly balance expected return and worst-case performance [47], while coherent distortion risk measures offer formal robustness guarantees in safe RL [34]. Other approaches focus on adversarial robustness or model-based safety. Some works combine robust model predictive control (MPC) with tube-based constraints to ensure recursive feasibility under uncertainty [51]. Gaussian Processes have also been used as safety oracles in model-based RL to probabilistically identify constraints [2]. Adversarially robust methods further address observation perturbations via state-adversarial MDPs and policy regularization [24, 25, 52].

Fuzzy Measures in MDPs. Fuzzy logic provides an interpretable framework for quantifying and managing uncertainty in complex systems. Zadeh's fuzzy sets theory [50] laid the foundation for uncertainty measures. Building on this, possibility theory [11] emerged as a significant fuzzy approach to uncertainty quantification, offering an alternative to probabilistic methods. Then, [26] introduced credibility theory, which combines fuzzy and probability measures to create a self-dual measure. For RL community, fuzzy Q-learning [13] and possibilistic MDPs [36] incorporate fuzzy logic into MDPs. Furthermore, [17], [21] and [19] developed various fuzzy RL approaches that provide enhanced interpretability and effectiveness compared to deep neural network-based RL methods. [15] introduces m_λ measure, which combined the possibility measure and necessity measure to balance optimism and pessimism in decision-making systems, which has shown promise in chance-constrained programming. Recent advancements have further expanded the application of fuzzy logic in uncertainty modeling. For instance, [20] developed a fuzzy adaptive sliding mode control method for robotic systems with uncertainties, [37] conducted a comprehensive review of uncertainty quantification applications for healthcare.

However, incorporating fuzzy logic for robustness enhancement in safe RL remains unexplored. Inspired by the fuzzy-guided m_λ fuzzy measure[15], we aim to achieve a robust risk-aversion and reward-pursuitin value estimation through fuzzy logic for safe RL.

3 Preliminaries

3.1 Robust CMDP

We consider formulating the safe RL problem as an infinite-horizon CMDP [3], which is defined by the tuple $(\mathcal{S}, \mathcal{A}, p, r, c, \gamma, d_0)$, where $\mathcal{S}$ is the finite state space and $\mathcal{A}$ is the action space. $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition model, $r, c : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ are the bounded reward function and cost function defining the objective and constraint, respectively. $\gamma \in [0, 1)$ is the discount factor, and $d_0 : \mathcal{S} \to [0, 1]$ is the initial state distribution. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps states to distributions over actions.

To address system uncertainty, we introduce a robust formulation of CMDP. Following the concept of (s, a)-rectangular uncertainty sets [32], we define the uncertainty set $\mathcal{P}$ as:

$$\mathcal{P} = \prod_{s,a} \mathcal{P}_s^a, \quad \mathcal{P}_s^a \subseteq \Delta(\mathcal{S}) \tag{1}$$

where $\mathcal{P}_s^a$ represents the set of all possible transition probabilities over $\mathcal{S}$ for a given state-action pair (s, a). For $\forall s \in \mathcal{S}, a \in \mathcal{A}$, we define:

$$\mathcal{P}_s^a = \{p(\cdot|s, a) : d(p(\cdot|s, a), p_0(\cdot|s, a)) \leq \epsilon\} \tag{2}$$

where p_0 is the nominal transition model, $d(\cdot, \cdot)$ is a distance metric, and $\epsilon > 0$ defines the uncertainty radius.

The objective of the robust CMDP is to find a policy π that solves the following constrained optimization problem:

$$\max_\pi \min_{p \in \mathcal{P}} \mathbb{E}_{\tau \sim (\pi, p)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \quad \text{s.t.} \max_{p \in \mathcal{P}} \mathbb{E}_{\tau \sim (\pi, p)} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \right] \leq B \tag{3}$$

where $\tau \sim (\pi, p)$ denotes trajectories sampled according to $s_0 \sim d_0$, $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim p(\cdot|s_t, a_t)$, and $B > 0$ is the safety budget constraint.

For computational tractability, we partition the uncertainty set $\mathcal{P}_s^a$ into K distinct levels:

$$\mathcal{P}_s^a = \bigcup_{k=1}^{K} \mathcal{P}_{s,k}^a, \quad \mathcal{P}_{s,k}^a = \{p(\cdot|s, a) : \epsilon_{k-1} < d(p(\cdot|s, a), p_0(\cdot|s, a)) \leq \epsilon_k\} \tag{4}$$

where $0 = \epsilon_0 < \epsilon_1 < ... < \epsilon_K = \epsilon$ defines a sequence of increasing uncertainty thresholds.

3.2 Fuzzy Measures Fundamentals

Traditional probability measures treat uncertainties in a purely additive manner, assuming independent effects from different uncertainties. However, in real control systems, as the distance from nominal dynamics increases, the compound effects of uncertainties often exhibit super-additive behavior. For example, when considering two uncertainty sources A and B, their joint impact on system performance may be greater than the sum of their individual effects:

$$m(A \cup B) > m(A) + m(B) \tag{5}$$

Moreover, as the system deviates further from the nominal model, the impact on performance and safety constraints typically grows non-linearly.

To capture such non-additive effects, we first introduce the concept of fuzzy measure, which provides an interpretable way to assess the impacts of uncertainty by assigning non-additive weights to combinations of uncertainty levels. The formal definition is as follows:

Definition 3.1 (*Fuzzy Measure* [30])**.** For each state-action pair (s, a), let $\mathcal{I}_s^a = \{\mathcal{P}_{s,1}^a, \mathcal{P}_{s,2}^a, \ldots, \mathcal{P}_{s,K}^a\}$ denote the collection of uncertainty sets. A fuzzy measure m is a set function $m : 2^{\mathcal{I}_s^a} \to [0, 1]$, satisfying:

(1) $m(\emptyset) = 0, \quad m(\mathcal{I}_s^a) = 1$ (boundary conditions),

(2) If $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{I}_s^a$, then $m(\mathcal{A}) \leq m(\mathcal{B})$ (monotonicity).

Measuring uncertainty impacts for all possible subset combinations in $2^{\{\mathcal{P}_{s,1}^a, \mathcal{P}_{s,2}^a, \ldots, \mathcal{P}_{s,K}^a\}}$ is computationally intractable, as it requires an exponential number of samples. To address this computational challenge while preserving the ability to model super-additive effects, we adopt the λ-fuzzy measure, which offers an efficient parameterization of subset relationships:

Definition 3.2 (λ-*Fuzzy Measure* [10])**.** A λ-fuzzy measure satisfies, for all disjoint subsets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{I}_s^a$ with $\mathcal{A} \cap \mathcal{B} = \emptyset$:

$$m(\mathcal{A} \cup \mathcal{B}) = m(\mathcal{A}) + m(\mathcal{B}) + \lambda\, m(\mathcal{A})\, m(\mathcal{B}), \tag{6}$$

where $\lambda > -1$ determines the degree of interaction. When $\lambda \in (-1, 0)$, the measure exhibits *sub-additive* behavior; when $\lambda \in (0, \infty)$, it models *super-additive* effects among different uncertainties. When $\lambda = 0$, the λ-fuzzy measure reduces to a normalized additive measure, i.e., a probability measure.

The connection between λ-fuzzy measures and robust optimization is established through the Choquet integral's pessimistic characterization:

Lemma 3.3 (*Choquet Integral Representation* [12])**.** *For any bounded measurable function $f : \mathcal{P}_s^a \to \mathbb{R}$ and convex λ-fuzzy measure m on $\mathcal{I}_s^a$ with $\lambda \geq 0$:*

$$(C) \int_{\mathcal{P}_s^a} f\, dm = \min_{P \in core(m)} \mathbb{E}_P[f],$$

where $core(m) = \{P : P(\mathcal{A}) \geq m(\mathcal{A}), \forall \mathcal{A} \subseteq \mathcal{I}_s^a\}$ is the set of probability measures on $\mathcal{P}_s^a$ dominating m.

4 Fuzzy Measure-based Robust Safe RL Framework

4.1 Theoretical Foundation of Fuz-RL

In this section, we connect fuzzy measure with robust CMDP by introducing the *Fuzzy Bellman operator*.

Fuzzy Bellman Operator. Leveraging Lemma 3.3, we define the fuzzy Bellman operator that encodes worst-case scenarios through the Choquet integral in Definition 4.1:

Definition 4.1 (*Fuzzy Bellman Operator*)**.** Let $\mathcal{B}(\mathcal{S})$ denote the space of bounded measurable value functions V on $\mathcal{S}$. The fuzzy Bellman operator $\mathcal{F} : \mathcal{B}(\mathcal{S}) \to \mathcal{B}(\mathcal{S})$ is defined as:

$$\mathcal{F}(V)(s) = \mathbb{E}_{a \sim \pi}\Big[r(s, a) + \gamma\, (C) \int_{\mathcal{P}_s^a} \mathbb{E}_{s' \sim p}[V(s')]\, dm(p)\Big].$$

where $m(\cdot)$ is the convex fuzzy measure.

Furthermore, we demonstrate that the fuzzy Bellman operator maintains fundamental properties of the standard Bellman operator (γ-contraction and convergence) when integrated with value functions as detailed in Theorem 4.2 and Theorem 4.3. Therefore, the fuzzy Bellman operator can be seamlessly integrated into value functions, enabling the establishment of robust value estimation with theoretical guarantees.

Theorem 4.2 (γ-*contraction of Fuzzy Bellman Operator*)**.** *For any $V_1, V_2 \in \mathcal{B}(\mathcal{S})$,*

$$\|\mathcal{F}(V_1) - \mathcal{F}(V_2)\|_\infty \leq \gamma \|V_1 - V_2\|_\infty.$$

Theorem 4.3 (*Convergence of Fuzzy Bellman Operator*)**.** *Let $V^0 \in \mathcal{B}(\mathcal{S})$ be an initial value function and $V^{n+1} = \mathcal{F}(V^n)$. Then V^n converges to a unique fixed point V^* satisfying $V^* = \mathcal{F}(V^*)$ with geometric rate $\|V^n - V^*\|_\infty \leq \gamma^n \|V^0 - V^*\|_\infty$.*

Robust Equivalence. Applying Lemma 3.3 shows that the Choquet integral automatically encodes a robust perspective via the fuzzy measure m. Consequently, we can prove the following Theorem 4.4:

Theorem 4.4 (*Equivalent Theorem*). *Let m be a convex λ-fuzzy measure on $\mathcal{I}_s^a$ defined by Definition 3.2 such that: (1) core$(m) \subseteq \mathcal{P}_s^a$ for all (s, a), (2) $\arg\min_{p \in \mathcal{P}_s^a} \mathbb{E}_{s' \sim p}[V(s')] \subseteq$ core(m) for all (s, a), (3) $\arg\max_{p \in \mathcal{P}_s^a} \mathbb{E}_{s' \sim p}[V_c(s')] \subseteq$ core(m) for all (s, a). Define the dual fuzzy measure $m'(\mathcal{A}) := 1 - m(\mathcal{I}_s^a \setminus \mathcal{A})$ for all $\mathcal{A} \subseteq \mathcal{I}_s^a$.*

*Then the **fuzzy robust safe RL problem (CMDP form)**:*

$$\max_{\pi} J_r^{\mathcal{F}}(\pi) \quad \text{s.t.} \quad J_c^{\mathcal{F}}(\pi) \leq B \tag{7}$$

where

$$J_r^{\mathcal{F}}(\pi) = (C) \int_{\mathcal{P}} \mathbb{E}_{\tau \sim (\pi, p)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] dm(p) \tag{8}$$

$$J_c^{\mathcal{F}}(\pi) = (C) \int_{\mathcal{P}} \mathbb{E}_{\tau \sim (\pi, p)} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \right] dm'(p) \tag{9}$$

is equivalent to the distributionally robust safe RL problem (robust CMDP form).

The detailed proofs of Theorem 4.2, Theorem 4.3 and Theorem 4.4 are provided in Appendix A.

4.2 Practical Implementation of Fuz-RL

Having established the fuzzy Bellman operator and its theoretical properties, we now describe how to implement Fuz-RL in practice. The pseudo-code of Fuz-RL is detailed in Appendix B Algorithm 1.

Estimation of Fuzzy Measures. To operationalize the fuzzy Bellman operator, we require an efficient method for estimating fuzzy measures $m(\cdot)$ that capture uncertainty impacts across different perturbation subsets. We adopt a neural network-based approach that learns fuzzy measure densities directly from state representations.

For each state s, we employ a fuzzy network ϕ parameterized as a multi-layer perceptron (MLP) with two hidden layers. The network takes the state vector as input and outputs fuzzy density parameters $\{g_k\}_{k=1}^K$:

$$\mathbf{g} = \sigma(\text{MLP}_{\phi}(s)), \tag{10}$$

where $\sigma(\cdot)$ denotes the softmax activation function applied to the network output. To maintain numerical stability, we apply clamping to constrain the fuzzy values within $[\epsilon, 1 - \epsilon]$ where $\epsilon = 10^{-4}$, preventing degenerate solutions while satisfying the mathematical constraints of λ-fuzzy measures.

Given the learned densities $\mathbf{g} = (g_1, \dots, g_K)$, the interaction parameter λ is determined by solving the characteristic equation:

$$\prod_{k=1}^{K} (1 + \lambda g_k) = 1 + \lambda, \tag{11}$$

using a hybrid bisection-Newton method with gradient detachment to ensure numerical stability. Once λ is obtained, the fuzzy measure for any index subset $A \subseteq \{1, \dots, K\}$ can be computed via the λ-rule:

$$m(A) = \frac{\prod_{k \in A}(1 + \lambda g_k) - 1}{\lambda}. \tag{12}$$

To simulate the impact of uncertain system state transitions across different uncertainty levels, we employ a stratified perturbation sampling strategy. For each uncertainty level $k \in \{1, \dots, K\}$, we apply independent isotropic Gaussian perturbations:

$$\tilde{s}_k = s + \epsilon_k \cdot \mathbf{n}_k, \quad \mathbf{n}_k \sim \mathcal{N}(0, I), \tag{13}$$

where ϵ_k represents the perturbation scale for uncertainty level k, typically set as $\epsilon_k = \epsilon_{\text{base}} \cdot k$ to create a hierarchy of perturbation intensities. For each uncertainty level, we generate $M = 5$ independent samples and compute $V(\tilde{s}_k)$ as the average of the value estimates across these samples to reduce estimation variance.

Estimation of Choquet Integrals. To approximate the Choquet integrals used in the fuzzy Bellman operator, we leverage the globally learned fuzzy measures. For the standard Choquet integral applied to reward value aggregation, we sort the perturbed value estimates in descending order: $V(\tilde{s}_{(1)}) \geq V(\tilde{s}_{(2)}) \geq \cdots \geq V(\tilde{s}_{(K)})$, where (i) denotes the index after sorting. The corresponding fuzzy measures are computed as $m_i = m(\{(i), (i+1), \ldots, (K)\})$, representing the capacity of the tail sets. Following the discrete Choquet integral formulation, the robust reward value is approximated as:

$$(C) \int_{\mathcal{P}_s^a} \mathop{\mathbb{E}}_{s' \sim p} [V(s')] \, dm(p) \approx \sum_{i=1}^{K} V(\tilde{s}_{(i)}) \big(m_i - m_{i+1} \big), \tag{14}$$

where $m_{K+1} = 0$ by convention. For the dual Choquet integral used in cost value aggregation with pessimistic estimation, we sort the cost values in ascending order: $V_c(\tilde{s}_{(1)}) \leq V_c(\tilde{s}_{(2)}) \leq \cdots \leq V_c(\tilde{s}_{(K)})$, and compute:

$$(C) \int_{\mathcal{P}_s^a} \mathop{\mathbb{E}}_{s' \sim p} [V_c(s')] \, dm'(p) \approx \sum_{i=1}^{K} V_c(\tilde{s}_{(i)}) \big(m_i - m_{i+1} \big), \tag{15}$$

where $m_i = m(\{1, 2, \ldots, i\})$ represents the capacity of the head sets, computed using the same global fuzzy measure densities g_k but with different subset selection to achieve pessimistic estimation for costs. The choice of descending order sorting for rewards and ascending order sorting for costs is theoretically grounded in the dual relationship between m and its dual measure m^*.

Value Network Updates. The value networks are updated through temporal difference learning with Choquet-integrated targets:

$$\begin{aligned} \mathcal{L}(\theta_r) &= \mathbb{E}_\tau \Big[\big(V_{\theta_r}(s_t) - \big(r_t + \gamma \cdot \hat{V}_{\theta_r}(s_{t+1}) \big) \big)^2 \Big], \\ \mathcal{L}(\theta_c) &= \mathbb{E}_\tau \Big[\big(V_{\theta_c}(s_t) - \big(c_t + \gamma \cdot \hat{V}_{\theta_c}(s_{t+1}) \big) \big)^2 \Big], \end{aligned} \tag{16}$$

where $\hat{V}_{\theta_r}$ and $\hat{V}_{\theta_c}$ denote the Choquet-integrated value estimates computed using Equations (14) and (15), respectively.

Fuzzy Network Updates. The fuzzy network parameters ϕ are updated at a lower frequency than the value networks to ensure stable learning of the uncertainty structure. Given the fuzzy density parameters $g(s_{t+1})$ predicted from next states, the network minimizes the discrepancy between Choquet-integrated predictions and Monte Carlo targets:

$$L(\phi) = \mathbb{E}_{\tau'} \Big[\big(r_t + \gamma \cdot V(s_{t+1}) - R_t \big)^2 + \big(c_t + \gamma \cdot V_c(s_{t+1}) - C_t \big)^2 \Big], \tag{17}$$

where R_t, C_t are Monte Carlo returns of reward r_t and cost c_t.

Policy Network Updates. Given that fuzzy value estimation implicitly addresses robustness through Choquet integration over multiple perturbations, the robust CMDP problem is solved using a primal-dual approach:

$$\max_{\pi} \min_{\lambda_\pi \geq 0} \mathbb{E}_{s \sim d^\pi} \big[V_r(s) - \lambda_\pi \big(V_c(s) - B \big) \big], \tag{18}$$

where V_r and V_c represent the robust value estimates obtained through Choquet integration. The policy parameters are optimized to maximize the Lagrangian objective, while the Lagrange multiplier is adjusted to enforce the safety constraint, with specific update rules determined by the underlying safe RL algorithm framework.

5 Experiments

5.1 Experiments Setup

To fully evaluate the robustness of Fuz-RL in multi-source uncertainties, we conduct experiments on four Safe-Control-Gym [6] tasks: `Cartpole-Stab`, `Cartpole-Track`, `Quadrotor-Stab`, and `Quadrotor-Track`, as well as four safety-critical control tasks with larger state-action spaces from Safety-Gymnasium [22]: `Safety-PointGoal1-v0`, `Safety-PointButton1-v0`, `Safety-PointCircle1-v0`, and `Safety-PointPush1-v0`.

Uncertainty Setting. During the test phase, we leverage different perturbations provided by the Safe-Control-Gym to consider the following settings on observation, action, and dynamics:

(1) *Observation uncertainty*. We introduce white noise following a normal distribution $\varepsilon \cdot \mathcal{N}(0, I)$ as observation perturbation, where ε is an adjustable parameter used to set different perturbation intensities. During testing, we vary ε from -1 to 1 in increments of 0.1 for Safe-Control-Gym tasks.

(2) *Action uncertainty.* We simulate action uncertainty through an impulse noise disturbance model. Specifically, the perturbed action $\bar{a}_t$ is formulated as $\bar{a}_t = a_t + d_t$, where d_t is defined as:

$$d_t = \begin{cases} \varepsilon M & t \in [t_{\text{start}}, t_{\text{start}} + D] \\ \varepsilon M \gamma^{(t - t_{\text{start}} - D)} & t > t_{\text{start}} + D \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

where $\varepsilon \in [-0.1, 0.1]$ is the magnitude coefficient, $M = 10$ is the amplification factor, $t_{start} = 20$ is the start step, $D = 80$ is the duration, and $\gamma = 0.9$ is the decay rate.

(3) *Dynamics uncertainty*. We apply white noise following a normal distribution $\varepsilon \cdot \mathcal{N}(0, I)$ to dynamics parameters—such as pole length and quadrotor mass—where the variation of ε follows the same scheme as that used for observation perturbation.

(4) *Multi-source uncertainty*. To analyze the non-additive nature of uncertainty perturbations, we simultaneously apply all three uncertainty settings in a coupled configuration during both training and testing. For Safety-Gymnasium tasks, we apply relatively small environmental perturbations during training with $\varepsilon = 0.5$. For Fuz-series algorithms, we uniformly adopt the training configuration with $\varepsilon_{\text{base}} = 0.1$, $K = 10$, and $M = 5$.

Since the Safety-Gymnasium benchmarks do not provide built-in uncertainty interfaces, we take observation uncertainty with $\varepsilon \in [0, 0.5]$ as an example for testing.

Baselines. We adopt three safe RL algorithms as the baseline, including the PPO-Lagrangian (PPOL) [35], Conservative Update Policy (CUP) [46], and CVaR-Proximal-Policy-Optimization (CPPO) [47]. After integrating the proposed fuzzy-guided framework, we obtain the corresponding Fuz-PPOL, Fuz-CUP, and Fuz-CPPO algorithms. Besides, the current state-of-the-art, robust safe RL, Risk-Averse Model Uncertainty (RAMU) [34], has also been migrated to our test tasks. All codes of Fuz-RL are implemented based on the SpinningUp [1].

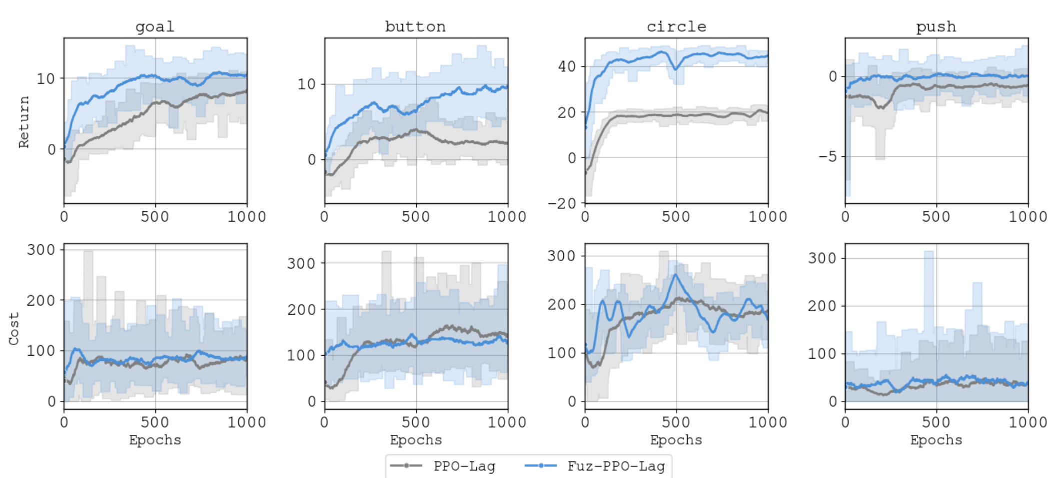

Figure 1: **Training Dynamics** of PPOLag and Fuz-PPOLag under `multi-source` uncertainty on Safety-Gymnasium tasks. The perturbation intensity during training is set to $\varepsilon = 0.5$.

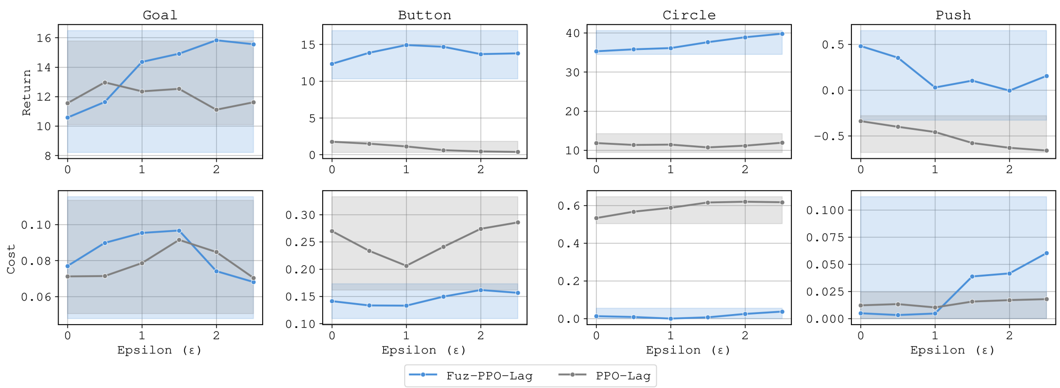

Figure 2: **Test Comparison** of PPOLag and Fuz-PPOLag under `multi-source` uncertainty setting over 5 Episodes and 5 seeds on Safety-Gymnasium tasks. The `cost_limit` is set to 0.1.

Table 1: **Detailed evaluation of Safe RL, Fuz-RL, and RAMU** on Safe-Control-Gym tasks with observation, action, dynamics uncertainty. Each value is reported as mean ± standard deviation for 10 episodes and 10 seeds. We shadow the highest AvgRet and lowest AvgRisk for each task.

Tasks	Methods	Observation Uncertainty		Action Uncertainty		Dynamics Uncertainty	
		AvgRet ↑	AvgRisk ↓	AvgRet ↑	AvgRisk ↓	AvgRet ↑	AvgRisk ↓
Cartpole Stab	PPOL	45 ± 28	0.40 ± 0.14	99 ± 19	0.27 ± 0.14	87 ± 16	0.27 ± 0.09
	CUP	32 ± 26	0.29 ± 0.11	42 ± 18	0.30 ± 0.13	50 ± 14	0.30 ± 0.13
	CPPO	41 ± 19	0.47 ± 0.10	76 ± 19	0.36 ± 0.14	77 ± 14	0.31 ± 0.11
	RAMU	35 ± 22	0.49 ± 0.16	86 ± 10	0.28 ± 0.07	86 ± 13	**0.20 ± 0.01**
	Fuz-PPOL	47 ± 25 ↑ 2	0.34 ± 0.11 ↓ 0.06	102 ± 17 ↑ 3	0.24 ± 0.12 ↓ 0.03	93 ± 13 ↑ 6	0.22 ± 0.08 ↓ 0.05
	Fuz-CUP	40 ± 19 ↑ 8	**0.26 ± 0.09** ↓ 0.03	87 ± 17 ↑ 45	**0.23 ± 0.13** ↓ 0.07	74 ± 13 ↑ 24	0.25 ± 0.08 ↓ 0.05
	Fuz-CPPO	**59 ± 25** ↑ 18	0.32 ± 0.10 ↓ 0.15	98 ± 17 ↑ 22	0.26 ± 0.13 ↓ 0.10	82 ± 14 ↑ 5	0.28 ± 0.09 ↓ 0.03
Cartpole Track	PPOL	70 ± 16	0.35 ± 0.14	95 ± 12	0.18 ± 0.10	91 ± 12	0.21 ± 0.10
	CUP	59 ± 12	0.28 ± 0.11	73 ± 9	0.20 ± 0.10	77 ± 11	0.18 ± 0.08
	CPPO	87 ± 17	0.42 ± 0.11	113 ± 16	0.26 ± 0.11	106 ± 12	0.32 ± 0.09
	RAMU	67 ± 31	**0.22 ± 0.13**	70 ± 34	0.18 ± 0.10	79 ± 31	0.14 ± 0.07
	Fuz-PPOL	91 ± 20 ↑ 21	0.38 ± 0.16 ↑ 0.03	**120 ± 14** ↑ 25	**0.18 ± 0.09** ↓ 0.00	**112 ± 14** ↑ 21	0.22 ± 0.09 ↑ 0.01
	Fuz-CUP	61 ± 31 ↑ 2	0.24 ± 0.16 ↓ 0.04	106 ± 16 ↑ 33	0.18 ± 0.11 ↓ 0.02	100 ± 12 ↑ 23	**0.14 ± 0.07** ↓ 0.04
	Fuz-CPPO	**93 ± 10** ↑ 6	0.31 ± 0.10 ↓ 0.11	107 ± 14 ↓ 6	0.21 ± 0.09 ↓ 0.05	107 ± 12 ↑ 1	0.23 ± 0.08 ↓ 0.09
Quadrotor Stab	PPOL	**164 ± 18**	0.13 ± 0.06	58 ± 55	0.52 ± 0.19	142 ± 34	0.28 ± 0.11
	CUP	139 ± 7	0.05 ± 0.02	58 ± 28	0.56 ± 0.17	117 ± 26	0.14 ± 0.10
	CPPO	131 ± 14	0.09 ± 0.02	54 ± 50	0.36 ± 0.11	117 ± 36	0.17 ± 0.13
	RAMU	146 ± 16	0.06 ± 0.04	28 ± 33	0.59 ± 0.20	120 ± 39	0.17 ± 0.16
	Fuz-PPOL	161 ± 22 ↓ 3	0.07 ± 0.05 ↓ 0.06	67 ± 58 ↑ 9	0.43 ± 0.19 ↓ 0.09	**156 ± 28** ↑ 14	0.13 ± 0.09 ↓ 0.15
	Fuz-CUP	142 ± 15 ↑ 3	**0.05 ± 0.02** ↓ 0.00	**94 ± 24** ↑ 36	**0.39 ± 0.10** ↓ 0.17	139 ± 23 ↑ 22	0.14 ± 0.09 ↓ 0.00
	Fuz-CPPO	156 ± 11 ↑ 25	0.07 ± 0.03 ↓ 0.02	87 ± 31 ↑ 33	0.33 ± 0.10 ↓ 0.03	130 ± 44 ↑ 13	**0.09 ± 0.12** ↓ 0.08
Quadrotor Track	PPOL	**218 ± 8**	0.48 ± 0.04	104 ± 79	0.81 ± 0.12	**203 ± 24**	0.58 ± 0.12
	CUP	151 ± 14	0.04 ± 0.03	67 ± 50	0.37 ± 0.13	152 ± 13	0.12 ± 0.11
	CPPO	152 ± 16	0.77 ± 0.04	76 ± 60	0.72 ± 0.11	124 ± 33	0.73 ± 0.08
	RAMU	176 ± 12	0.05 ± 0.03	61 ± 50	0.53 ± 0.18	123 ± 48	0.31 ± 0.20
	Fuz-PPOL	200 ± 6 ↓ 18	0.28 ± 0.05 ↓ 0.20	99 ± 67 ↓ 5	0.64 ± 0.17 ↓ 0.17	194 ± 16 ↓ 9	0.38 ± 0.15 ↓ 0.20
	Fuz-CUP	175 ± 9 ↑ 24	**0.04 ± 0.02** ↓ 0.00	**112 ± 22** ↑ 45	**0.33 ± 0.09** ↓ 0.04	168 ± 13 ↑ 16	**0.12 ± 0.10** ↓ 0.00
	Fuz-CPPO	168 ± 14 ↑ 16	0.47 ± 0.05 ↓ 0.30	79 ± 53 ↑ 3	0.67 ± 0.09 ↓ 0.05	151 ± 23 ↑ 27	0.59 ± 0.12 ↓ 0.14

5.2 Robustness Assessment in Safe Control Tasks

For Safe-Control-Gym tasks, we trained the three safe RL baseline algorithms along with their corresponding Fuz-RL and RAMU variants under the same configuration. The specific parameter settings and more detailed results are presented in Appendix C.2. For Safety-Gymnasium tasks, we use PPOLag and Fuz-PPOLag as representative examples for experimental evaluation, with training and testing dynamics shown in Figure 1 and Figure 2, respectively.

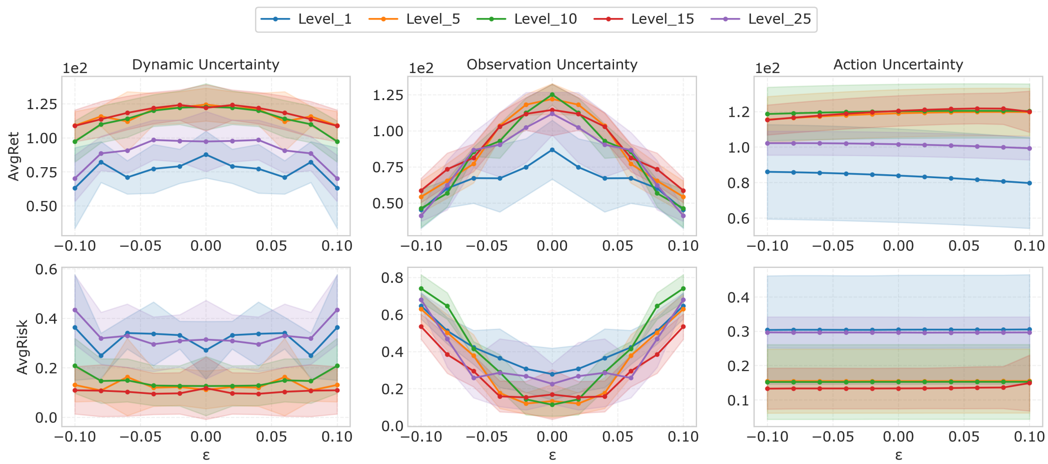

Figure 3: Ablation study of the uncertainty level K.

We evaluate the models from two features "*AvgRet*" and "*AvgRisk*", which represent the average episodic return and the *proportion of constraint violations*, respectively, for each task over 10 episodes across 10 seeds.

Comparison between Safe RL and Fuz-RL. As depicted in Table (1) and Appendix C.3, Fuz-RL demonstrates superior safety in **94.9%** cases and robust performance in **88.9%** tasks across various uncertainty settings. Taking Fuz-CUP as an example, it achieves 61.4% higher *AvgRet* and 16.7% lower *AvgRisk* than CUP in CartPole-Stab task. Moreover, Fuz-RL shows better uncertainty resistance with slower performance degradation than Safe RL. The variance reduction in *AvgRet* is 20.7%, 9.9%, and 8.6%, while in *AvgRisk* is 13.2%, 7.1%, and 22.6% respectively.

Comparison between Fuz-RL and RAMU. In the 36 Fuz-RL-based experiments listed in Table (1), Fuz-RL surpasses RAMU in achieving higher *AvgRet* in **83.3%** of the tasks. Furthermore, Fuz-RL exhibits lower *AvgRisk* compared to RAMU in 52.8% of them. It is important to highlight that *the lower average episodic risk* of RAMU is achieved by *compromising average episodic rewards*, especially in cases of actions affected by impulse disturbances, as shown in the "Action Uncertainty" section of Table (1). Fuz-RL consistently outperforms RAMU in all "Action Uncertainty" tasks.

Ablation Studies of Fuz-RL. We conduct ablation studies to examine how different uncertainty levels K affect Fuz-RL's performance. As illustrated in Figure 3, setting uncertainty level ($K = 1$) proves insufficient and leads to high episodic risk, while excessive levels ($K = 25$) complicate training and reduce rewards. The optimal performance emerges at intermediate levels ($K = 5$ to $K = 15$), where agents achieve higher rewards while maintaining lower and more stable risk across all three uncertainty types.

6 Conclusion and Future Work

In this paper, we propose Fuz-RL, a novel robustness enhancement framework that seamlessly integrates fuzzy logic into safe reinforcement learning. We develop a novel fuzzy Bellman operator incorporating Choquet integrals, enabling robust decision-making without solving computationally expensive min-max optimization problems. Theoretically, we establish the equivalence between our fuzzy robust safe RL formulation and distributionally robust safe RL. Extensive experiments on the safe-control-gym and safety-gymnasium benchmarks demonstrate that Fuz-RL significantly outperforms state-of-the-art safe and robust RL algorithms across various uncertainty types, achieving superior performance in both reward optimization and safety constraint satisfaction under diverse perturbation scenarios.

While Fuz-RL demonstrates promising results, it faces limited scalability in high-dimensional state spaces. Future work will focus on developing more efficient uncertainty modeling techniques and extending the framework to handle non-stationary uncertainty distributions through adaptive learning mechanisms.

Acknowledgements

This work was supported in part by the Smart Grid-National Science and Technology Major Project (2025ZD0803600, 2025ZD0803604), the National Natural Science Foundation of China under Grants 72571007, and the Natural Science Foundation of Zhejiang Province under Grant LZ23F030009.

References

[1] Joshua Achiam. Spinning up in deep reinforcement learning.(2018). *URL https://github.com/openai/spinningup*, 2018.

[2] Filippo Airaldi, Bart De Schutter, and Azita Dabiri. Learning safety in model-based reinforcement learning using mpc and gaussian processes. *IFAC-PapersOnLine*, 56(2):5759–5764, 2023.

[3] Eitan Altman. *Constrained Markov decision processes: stochastic modeling*. Routledge, 1999.

[4] Aaron D Ames, Xiangru Xu, Jessy W Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2016.

[5] Yarden As, Ilnura Usmanova, Sebastian Curi, and Andreas Krause. Constrained policy optimization via bayesian world models. *arXiv preprint arXiv:2201.09802*, 2022.

[6] Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 2021.

[7] Hyo-Nam Cho, Hyun-Ho Choi, and Yoon-Bae Kim. A risk assessment methodology for incorporating uncertainties using fuzzy concepts. *Reliability Engineering & System Safety*, 78(2):173–183, 2002.

[8] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(167):1–51, 2018.

[9] Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

[10] Dieter Denneberg. *Non-additive measure and integral*, volume 27. Springer Science & Business Media, 1994.

[11] Didier Dubois and Henri Prade. *Possibility theory: an approach to computerized processing of uncertainty*. Plenum Press, New York, 1988.

[12] Itzhak Gilboa and David Schmeidler. Additive representations of non-additive measures and the choquet integral. *Annals of Operations Research*, 52:43–65, 1994.

[13] Pierre Yves Glorennec and Lionel Jouffe. Fuzzy q-learning. In *Proceedings of 6th international fuzzy systems conference*, volume 2, pages 659–662. IEEE, 1997.

[14] Michel Grabisch et al. *Set functions, games and capacities in decision making*, volume 46. Springer, 2016.

[15] Peng Guo and Yue Wang. Fuzzy chance-constrained programming with linear combination of possibility measure and necessity measure. *Applied Mathematical Modelling*, 49:562–578, 2018.

[16] Yusuf Güven, Ata Köklü, and Tufan Kumbasar. Exploring zadeh's general type-2 fuzzy logic systems for uncertainty quantification. *IEEE Transactions on Fuzzy Systems*, 2024.

[17] Daniel Hein, Alexander Hentschel, Thomas Runkler, and Steffen Udluft. Particle swarm optimization for generating interpretable fuzzy reinforcement learning policies. *Engineering Applications of Artificial Intelligence*, 65:87–98, 2017.

[18] Abdelfetah Hentout, Abderraouf Maoudj, and Mustapha Aouache. A review of the literature on fuzzy-logic approaches for collision-free path planning of manipulator robots. *Artificial Intelligence Review*, 56(4):3369–3444, 2023.

[19] Jordan Hostetter, Mohamed Abdelshiheed, Tiffany Barnes, and Min Chi. A self-organizing neuro-fuzzy q-network: Systematic design with offline hybrid learning. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2023.

[20] Changchun Hua, Shaocheng Tong, Yongming Li, and Jun Cheng. Fuzzy adaptive sliding mode control for robotic systems with uncertainties. *IEEE Transactions on Fuzzy Systems*, 31(6):1849–1859, 2023.

[21] Jianyu Huang, Plamen P Angelov, and Chengbin Yin. Interpretable policies for reinforcement learning by empirical fuzzy sets. *Engineering Applications of Artificial Intelligence*, 91:103559, 2020.

[22] Jiaming Ji, Borong Zhang, Jiayi Zhou, Xuehai Pan, Weidong Huang, Ruiyang Sun, Yiran Geng, Yifan Zhong, Josef Dai, and Yaodong Yang. Safety gymnasium: A unified safe reinforcement learning benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[23] Ezgi Korkmaz. Deep reinforcement learning policies learn shared adversarial features across mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7229–7238, 2022.

[24] Zeyang Li, Chuxiong Hu, Yunan Wang, Yujie Yang, and Shengbo Eben Li. Safe reinforcement learning with dual robustness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[25] Yongyuan Liang, Yanchao Sun, Ruijie Zheng, and Furong Huang. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. *Advances in neural information processing systems*, 35:22547–22561, 2022.

[26] Yong-Jin Liu and Baoding Liu. *Fuzzy Radon and Choquet integrals*. Fuzzy Sets and Systems, 2002.

[27] Zhi Liu, Fang Wang, Yun Zhang, and CL Philip Chen. Fuzzy adaptive quantized control for a class of stochastic nonlinear uncertain systems. *IEEE transactions on cybernetics*, 46(2):524–534, 2015.

[28] Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust q-learning. In *International Conference on Machine Learning*, pages 13623–13643. PMLR, 2022.

[29] Jun Morimoto and Kenji Doya. Robust reinforcement learning. *Neural computation*, 17(2):335–359, 2005.

[30] Toshiaki Murofushi, Michio Sugeno, et al. Fuzzy measures and fuzzy integrals. *Fuzzy measures and integrals: theory and applications*, 2000:3–41, 2000.

[31] Xinyi Ni and Lifeng Lai. Risk-sensitive reinforcement learning via entropic-var optimization. In *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pages 953–959. IEEE, 2022.

[32] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[33] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR, 2017.

[34] James Queeney and Mouhacine Benosman. Risk-averse model uncertainty for distributionally robust safe reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[35] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019.

[36] Régis Sabbadin. Possibilistic markov decision processes. *Engineering Applications of Artificial Intelligence*, 14(3):287–300, 2001.

[37] Silvia Seoni, Vicnesh Jahmunah, Massimo Salvi, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Artificial Intelligence in Medicine*, 140:102569, 2023.

[38] Rahul Singh, Qinsheng Zhang, and Yongxin Chen. Improving robustness via risk averse distributional reinforcement learning. In *Learning for Dynamics and Control*, pages 958–968. PMLR, 2020.

[39] Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.

[40] Yoki Tanabe, Takayuki Yamaguchi, and Yasuhiro Shirai. Max-min reinforcement learning for robust policies. *IEEE Access*, 10:105370–105380, 2022.

[41] Chen Tessler, Yonathan Efroni, and Shie Mannor. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pages 6215–6224. PMLR, 2019.

[42] Akifumi Wachi, Wataru Hashimoto, Xun Shen, and Kazumune Hashimoto. Safe exploration in reinforcement learning: A generalized formulation and algorithms. *Advances in Neural Information Processing Systems*, 36, 2024.

[43] Xu Wan and Mingyang Sun. Adapsafe2: Prior-free safe-certified reinforcement learning for multi-area frequency control. *IEEE Transactions on Power Systems*, 2024.

[44] Xu Wan, Mingyang Sun, Boli Chen, Zhongda Chu, and Fei Teng. Adapsafe: adaptive and safe-certified deep reinforcement learning-based frequency control for carbon-neutral power systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5294–5302, 2023.

[45] Yixuan Wang, Simon Sinong Zhan, Ruochen Jiao, Zhilu Wang, Wanxin Jin, Zhuoran Yang, Zhaoran Wang, Chao Huang, and Qi Zhu. Enforcing hard constraints with soft barriers: Safe reinforcement learning in unknown stochastic environments. In *International Conference on Machine Learning*, pages 36593–36604. PMLR, 2023.

[46] Long Yang, Jiaming Ji, Juntao Dai, Yu Zhang, Pengfei Li, and Gang Pan. Cup: A conservative update policy algorithm for safe reinforcement learning. *arXiv preprint arXiv:2202.07565*, 2022.

[47] Chengyang Ying, Xinning Zhou, Dong Yan, and Jun Zhu. Towards safe reinforcement learning via constraining conditional value at risk. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

[48] Dongjie Yu, Haitong Ma, Shengbo Li, and Jianyu Chen. Reachability constrained reinforcement learning. In *International Conference on Machine Learning*, pages 25636–25655. PMLR, 2022.

[49] Jiong Yu and Athina P Petropulu. On max-min fair congestion control. *IEEE Communications Letters*, 26(11):1696–1698, 1996.

[50] Lotfi A Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[51] Mario Zanon and Sébastien Gros. Safe reinforcement learning using robust mpc. *IEEE Transactions on Automatic Control*, 66(8):3638–3652, 2020.

[52] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in neural information processing systems*, 33:21024–21037, 2020.

[53] Tao Zhao, Haodong Li, and Songyi Dian. Multi-robot path planning based on improved artificial potential field and fuzzy inference system. *Journal of Intelligent & Fuzzy Systems*, 39(5):7621–7637, 2020.

A Appendix / Theorems and Proofs

Theorem A.1 (γ-contraction of Fuzzy Bellman Operator)**.** *For any $V_1, V_2 \in \mathcal{B}(\mathcal{S})$,*

$$\|\mathcal{F}(V_1) - \mathcal{F}(V_2)\|_\infty \leq \gamma \|V_1 - V_2\|_\infty.$$

Proof. For any two value functions V_1 and V_2, and any state s:

$$\begin{aligned}
&|\mathcal{F}(V_1)(s) - \mathcal{F}(V_2)(s)| \\
&= \left| \mathbb{E}_{a\sim\pi} \left[\gamma (C) \int_{\mathcal{P}_s^a} \mathbb{E}_{s'\sim p}[V_1(s') - V_2(s')] dm(p) \right] \right| \\
&\leq \gamma \mathbb{E}_{a\sim\pi} \left[(C) \int_{\mathcal{P}_s^a} \mathbb{E}_{s'\sim p} |V_1(s') - V_2(s')| dm(p) \right] \\
&\leq \gamma \mathbb{E}_{a\sim\pi} \left[\|V_1 - V_2\|_\infty (C) \int_{\mathcal{P}_s^a} dm(p) \right] \\
&= \gamma \|V_1 - V_2\|_\infty \mathbb{E}_{a\sim\pi} \left[(C) \int_{\mathcal{P}_s^a} dm(p) \right] \\
&= \gamma \|V_1 - V_2\|_\infty
\end{aligned}$$

Here, we use the fact that $(C) \int_{\mathcal{P}_s^a} dm(p) = 1$ for all s and a, as m is a normalized fuzzy measure. Taking the supremum over all states s yields the result. □

Theorem A.2 (Convergence of Fuzzy Bellman Operator)**.** *Let $V^0 \in \mathcal{B}(\mathcal{S})$ be an initial value function and $V^{n+1} = \mathcal{F}(V^n)$. Then V^n converges to a unique fixed point V^* satisfying $V^* = \mathcal{F}(V^*)$ with geometric rate $\|V^n - V^*\|_\infty \leq \gamma^n \|V^0 - V^*\|_\infty$.*

Proof. By Theorem 4.2, $\mathcal{F}$ is a contraction mapping. The Banach Fixed Point Theorem guarantees the existence of a unique fixed point V^* such that $V^* = \mathcal{F}(V^*)$. Moreover, for any initial V^0, the sequence $\{V^n\}_{n=0}^\infty$ defined by $V^{n+1} = \mathcal{F}(V^n)$ converges to V^*:

$$\|V^n - V^*\|_\infty \leq \gamma^n \|V^0 - V^*\|_\infty \to 0 \text{ as } n \to \infty$$

This convergence follows directly from the contraction property:

$$\begin{aligned}
\|V^{n+1} - V^*\|_\infty &= \|\mathcal{F}(V^n) - \mathcal{F}(V^*)\|_\infty \\
&\leq \gamma \|V^n - V^*\|_\infty \\
&\leq \gamma^n \|V^1 - V^*\|_\infty \\
&\leq \gamma^n \|V^1 - V^0\|_\infty + \gamma^n \|V^0 - V^*\|_\infty
\end{aligned}$$

As $n \to \infty$, both terms approach zero due to $\gamma < 1$, proving the convergence. □

Lemma A.3 (Core Duality of Convex Fuzzy Measures)**.** *Let m be a convex fuzzy measure on $\mathcal{I}_s^a = \{\mathcal{P}_{s,1}^a, \ldots, \mathcal{P}_{s,K}^a\}$ with dual measure defined by:*

$$m'(\mathcal{A}) := 1 - m(\mathcal{I}_s^a \setminus \mathcal{A}), \quad \forall \mathcal{A} \subseteq \mathcal{I}_s^a.$$

Then the cores satisfy $\text{core}(m') = \text{core}(m)$*, and for any bounded measurable function $f : \mathcal{P}_s^a \to \mathbb{R}$:*

$$(C) \int_{\mathcal{P}_s^a} f \, dm' = \max_{P \in \text{core}(m)} \mathbb{E}_P[f].$$

Proof. **Part 1: Core Equivalence**. For any convex fuzzy measure m on $\mathcal{I}_s^a$, its dual m' is concave. By the duality theorem for balanced fuzzy measures [14], for any probability measure P on $\mathcal{P}_s^a$:

$$\begin{aligned}
P \in \text{core}(m) &\iff P(\bigcup_{\mathcal{P}\in\mathcal{A}} \mathcal{P}) \geq m(\mathcal{A}),\ \forall \mathcal{A} \subseteq \mathcal{I}_s^a \\
&\iff 1 - P(\bigcup_{\mathcal{P}\in\mathcal{I}_s^a \setminus \mathcal{A}} \mathcal{P}) \geq 1 - m(\mathcal{I}_s^a \setminus \mathcal{A}),\ \forall \mathcal{A} \subseteq \mathcal{I}_s^a \\
&\iff P(\bigcup_{\mathcal{P}\in\mathcal{A}} \mathcal{P}) \geq m'(\mathcal{A}),\ \forall \mathcal{A} \subseteq \mathcal{I}_s^a \\
&\iff P \in \text{core}(m').
\end{aligned}$$

Part 2: Maximum Representation. For the concave measure m', the Choquet integral attains its maximum over the core:

$$(C)\int f\, dm' = \max_{P\in\text{core}(m')} \mathbb{E}_P[f] = \max_{P\in\text{core}(m)} \mathbb{E}_P[f],$$

where the last equality follows from $\text{core}(m') = \text{core}(m)$. □

Theorem A.4 (Equivalence Theorem)**.** *Given a robust CMDP:*

$$\begin{aligned}
&\max_{\pi} \min_{p\in\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \\
\textit{s.t.} \quad &\max_{p\in\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right] \leq B,
\end{aligned}$$

let m be a convex λ-fuzzy measure on $\mathcal{I}_s^a$ such that:

1. *$\text{core}(m) \subseteq \mathcal{P}_s^a$ for all (s, a),*
2. *$\arg\min_{p\in\mathcal{P}_s^a} \mathbb{E}_{s'\sim p}[V(s')] \subseteq \text{core}(m)$ for all (s, a),*
3. *$\arg\max_{p\in\mathcal{P}_s^a} \mathbb{E}_{s'\sim p}[V_c(s')] \subseteq \text{core}(m)$ for all (s, a).*

Define the dual fuzzy measure $m'(\mathcal{A}) := 1 - m(\mathcal{I}_s^a \setminus \mathcal{A})$ for all $\mathcal{A} \subseteq \mathcal{I}_s^a$.

Then the Fuz-RL problem:

$$\max_{\pi} J_r^{\mathcal{F}}(\pi) \quad \textit{s.t.} \quad J_c^{\mathcal{F}}(\pi) \leq B,$$

where

$$J_r^{\mathcal{F}}(\pi) = (C)\int_{\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] dm(p) \tag{20}$$

$$J_c^{\mathcal{F}}(\pi) = (C)\int_{\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)} \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)\right] dm'(p) \tag{21}$$

is equivalent to the original robust CMDP.

Proof. **Step 1: Core Inclusion and Extremal Coverage**. By Equation 10 with softmax activation and clamping to $[\epsilon, 1-\epsilon]$ where $\epsilon = 10^{-4}$, the λ-fuzzy measure satisfies $m(\{\mathcal{P}_{s,k}^a\}) = g_k \in (0, 1)$ for each uncertainty level $k \in \{1, \ldots, K\}$ via Equation 12.

Since each uncertainty level $\mathcal{P}_{s,k}^a$ covers an ϵ_k-neighborhood of the nominal dynamics (Equation 4), and the softmax ensures all levels receive positive weights, extremal perturbations are guaranteed to belong to $\text{core}(m)$. This establishes conditions (2) and (3), ensuring $\text{core}(m) \subseteq \mathcal{P}_s^a$ and covering extremal points of $\mathcal{P}_s^a$.

Step 2: Duality of Fuzzy Measures. For the convex fuzzy measure m, its dual m' is concave. By Choquet duality [14]:

$$(C)\int f\,dm = \inf_{q\in \text{core}(m)} \mathbb{E}_q[f], \quad (C)\int f\,dm' = \sup_{q\in \text{core}(m)} \mathbb{E}_q[f].$$

Step 3: Reward Objective Equivalence. For the reward function, by Lemma 3.3:

$$\begin{aligned} J_r^{\mathcal{F}}(\pi) &= (C)\int_{\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] dm(p) \\ &\overset{(a)}{=} \min_{q\in \text{core}(m)} \mathbb{E}_{\tau\sim(\pi,q)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] \\ &\overset{(b)}{=} \min_{p\in\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right], \end{aligned}$$

where (a) uses the Choquet integral representation for convex measure m (Lemma 3.3), and (b) holds because condition (2) ensures that core(m) contains the extremal point $\arg\min_{p\in\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)}[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)]$.

Step 4: Cost Constraint Equivalence. For the cost function, by Lemma A.3:

$$\begin{aligned} J_c^{\mathcal{F}}(\pi) &= (C)\int_{\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right] dm'(p) \\ &\overset{(c)}{=} \max_{q\in \text{core}(m)} \mathbb{E}_{\tau\sim(\pi,q)}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right] \\ &\overset{(d)}{=} \max_{p\in\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right], \end{aligned}$$

where (c) uses the Choquet integral representation for concave measure m' (Lemma A.3), and (d) holds because condition (3) ensures that core(m) contains the extremal point $\arg\max_{p\in\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)}[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)]$.

Step 5: Equivalence Conclusion. Combining Steps 3–4, the Fuz-RL problem:

$$\max_{\pi} J_r^{\mathcal{F}}(\pi) \quad \text{s.t.} \quad J_c^{\mathcal{F}}(\pi) \le B$$

is equivalent to the original robust CMDP:

$$\begin{aligned} &\max_{\pi}\min_{p\in\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right] \\ \text{s.t.}\quad &\max_{p\in\mathcal{P}} \mathbb{E}_{\tau\sim(\pi,p)}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t)\right] \le B, \end{aligned}$$

as both objectives and constraints encode the same worst-case expectations over $\mathcal{P}$. □

B Appendix/Algorithm

B.1 Optimization Details

For optimal policy optimization based on value estimation, we solve the constrained optimization problem using the Lagrangian method:

$$\mathcal{L}(\pi,\lambda) = J_r^{\mathcal{F}}(\pi) - \lambda(J_c^{\mathcal{F}}(\pi) - B) \tag{22}$$

Algorithm 1 Fuzzy-Guided Robust Framework for Safe RL (Fuz-RL)

Input: actor θ_π, critics θ_r, θ_c, fuzzy density parameters g, uncertainty levels K, replay buffer $\mathcal{D}$.
Initialize: $\theta_r, \theta_c, \theta_\pi, g$, buffer $\mathcal{D}$.
for epoch $= 1$ to MaxEpoch **do**
  for $t = 1$ to T **do**
    Sample action $a_t \sim \pi_{\theta_\pi}(\cdot|s_t)$, observe s_{t+1} and get r_t, c_t.
    For each $i \in \{1, \ldots, K\}$, generate perturbed state $\tilde{s}_i = s + \epsilon_i \cdot \mathcal{N}(0, I)$.
    Store the tuple $(s_t, a_t, r_t, c_t, s_{t+1}, \{\tilde{s}_i\}_{i=1}^{K})$ in $\mathcal{D}$.
  end for
  for each *actor/critic network update step* **do**
    Sample mini-batch τ from $\mathcal{D}$.
    Compute perturbed values $\{V_{\theta_r}(\tilde{s}_i)\}$ and $\{V_{\theta_c}(\tilde{s}_i)\}$.
    Calculate Choquet integrals using Eqs. (14)–(15).
    Update $V_{\theta_r}, V_{\theta_c}$, fuzzy density parameters g using Eq. (16) and Eq. (17).
    Update π_{θ_π} using Eq. (18) with specific safe RL algorithm.
  end for
end for
Output: Trained parameters $\theta_\pi, \theta_r, \theta_c, g$.

The optimal policy π^* and the optimal Lagrangian multiplier λ^* can be obtained by:

$$(\pi^*, \lambda^*) = \arg\max_{\pi} \min_{\lambda \geq 0} \mathcal{L}(\pi, \lambda) \tag{23}$$

For a given state s, the optimal action selection rule becomes:

$$\pi^*(a|s) = \arg\max_{a \in \mathcal{A}} Q_r^\pi(s, a) - \lambda^* Q_c^\pi(s, a) \tag{24}$$

where the action-value functions are:

$$Q_r^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim p}[V_r^*(s')] \tag{25}$$

$$Q_c^\pi(s, a) = c(s, a) + \gamma \mathbb{E}_{s' \sim p}[V_c^*(s')] \tag{26}$$

Here, V_r^* and V_c^* are the unique fixed points guaranteed by Theorems 4.2 and 4.3, representing the optimal robust value functions for reward and cost, respectively.

In practice, we compute the optimal policy iteratively by initializing a Lagrangian multiplier $\lambda^{(0)}$ and then alternating between policy updates and multiplier updates. At each iteration k, we compute:

$$\pi^{(k)} = \arg\max_{\pi} J_r^{\mathcal{F}}(\pi) - \lambda^{(k)}(J_c^{\mathcal{F}}(\pi) - B) \tag{27}$$

$$\lambda^{(k+1)} = [\lambda^{(k)} + \alpha(J_c^{\mathcal{F}}(\pi^{(k)}) - B)]^+ \tag{28}$$

where $\alpha > 0$ is a step size and $[x]^+ = \max(0, x)$. This process continues until convergence, yielding the optimal safe policy π^* that maximizes reward while satisfying the safety constraint.

C Appendix / Experiment Setting and More Results

C.1 Environment description

C.1.1 Double Integrator

The following dynamics describe the double integrator:

$$\begin{bmatrix} x_{t+1} \\ v_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_t \\ v_t \end{bmatrix} + \begin{bmatrix} 0.005 \\ 0 \end{bmatrix} a_t, \tag{29}$$

where $a_t \in [-1, 1]$. The safety constraints are $|x| \leq 2$ and $|v| \leq 2$.

The reward function induced to unsafe state is designed as follows:

$$
\begin{aligned}
r(x,v) = &\max(4-(2(x-1.5)^2+2(v+1.5)^2),0)+ \\
&\max(5-(3(x+2.2)^2+3(v+2.2)^2),0)+ \\
&\max(5-(3(x-2.2)^2+3(v-2.2)^2),0)+ \\
&\max(4-(2(x+1.5)^2+2(v-1.5)^2),0)
\end{aligned} \tag{30}
$$

C.1.2 Safe Control Gym

The safe-control-gym benchmark comprises three dynamical systems: the Cartpole, and the 1D and 2D Quadrotors, as shown in Figure. 4. In our setting, we use CartPole and 2D QuadRotor as the base environments.

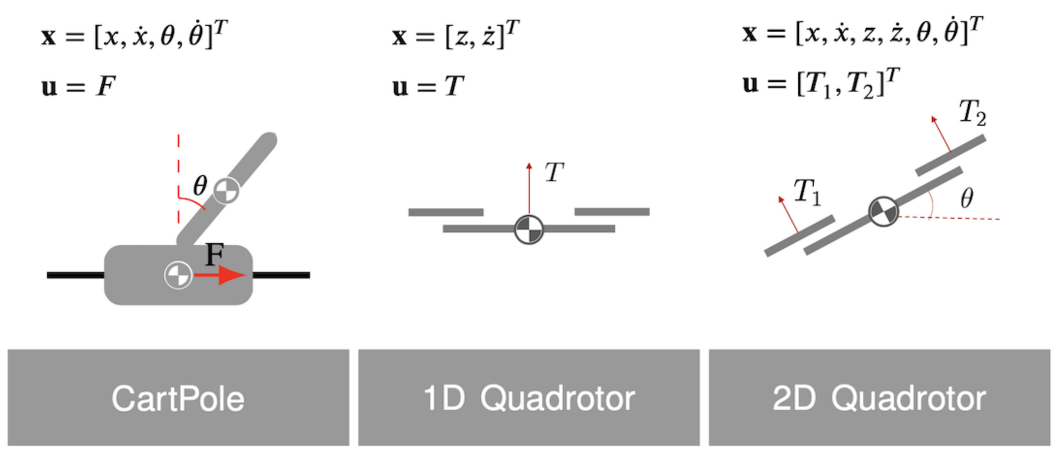

Figure 4: Schematics, state and input vectors of the cart-pole, and the 1D and 2D quadrotor environments in safe-control-gym.

For the CartPole system, the system state includes position x and velocity v of the cart, angle θ, and angular velocity ω of the pole. The control inputs $u \in [-1,1] \subset \mathbb{R}$ and external disturbances $a \in [-0.5, 0.5] \subset \mathbb{R}$ are horizontal forces applied on the cart. The safety constraints are $|\theta| \leq 0.2$, i.e., keeping the pole nearly upright. The constraint function is $h(\theta) = \min\{\theta + 0.2, 0.2 - \theta\}$.

For the 2D QuadRotor system, the state of the system is given by $\mathbf{s} = \left[x, \dot{x}, z, \dot{z}, \theta, \dot{\theta}\right]^T$, where (x, z) and $(\dot{x}, \dot{z})$ are the translation position and velocity of the COM of the quadrotor in the xz-plane, and θ and $\dot{\theta}$ are the pitch angle and the pitch angle rate, respectively. The input of the system is the thrusts $\mathbf{a} = [T_1, T_2]^T$ generated by two pairs of motors, one on each side of the body's y-axis. The safety constraints are $z - 0.5 > 0$ and $1.5 - z > 0$, i.e., maintaining its vertical position z between $[-0.5, 1.5]$. The constraint function is $h(z) = \min\{z - 0.5, 1.5 - z\}$.

For the reward function setup, we utilize a weighted sum of the errors between the current state $\mathbf{s}$, action $\mathbf{a}$, and their reference values as the reward for each step. The details of the weighting are provided in Table 2.

Besides, each environment in Safe-Control-Gym supports two control tasks: stabilization and trajectory tracking. For stabilization, safe-controlgym provides an equilibrium pair for the system, $x^{\text{ref}}, u^{\text{ref}}$. For trajectory tracking, the benchmark includes a trajectory generation module capable of generating circular, sinusoidal, lemniscate, or square trajectories. The module returns references $x_{\text{ref}_i},\ u_{\text{ref}_i}\ \forall i \in \{0, \ldots, L\}$, where L is the number of control steps in an episode.

C.2 Hyper-parameters

C.2.1 Hyper-parameters of RL

In all the experiments, we have revised the benchmark algorithms and Fuz-RL employing the RL framework provided by Spinning Up. The complete hyperparameters used in the experiments are shown in Table 2.

Particularly, for the CPPO and Fuz-CPPO algorithms, the risk threshold β for adverse trajectories is set to 100. In the case of the PPOL and CUP algorithms, the initial value of the Lagrange coefficient is set to 0.001, with an upper limit of 0.2 and a learning rate of 0.02. For the RAMU algorithm, the Wang transform is utilized with $\eta = 0.75$, which is applied to both the objective and the constraint.

Table 2: Hyperparameter Settings of Fuz-RL Training and Testing

	CartPole-Stab	CartPole-Track	QuadRotor-Stab	QuadRotor-Track
rollout length	150	150	250	250
training epoch	500	500	1000	1000
batch size	64	64	64	128
cost limit	1	1	10	10
uncertainty level K	10	10	15	15
optimization step	40	40	80	80
actor learning rate	0.0003	0.0003	0.0002	0.0002
critic learning rate	0.001	0.001	0.001	0.001
fuzzy learning rate	0.0003	0.0003	0.0003	0.0003
target KL	0.2	0.2	0.15	0.15
hidden_sizes	[64, 64]	[64, 64]	[256, 128]	[256, 128]
rew_act_weight	0.1	0.01	0.1	0.01
rew_state_weight	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.01 \\ 0.01 & 0.01 \end{bmatrix}$	$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.01 \\ 1 & 0.01 \\ 0.01 & 0.01 \end{bmatrix}$

Table 3: The observation, dynamics and action uncertainty settings of Safe-Control-Gym tasks

Uncertainty Object	Type	Config	System	
			CartPole	QuadRotor
Observation	white noise	std:[-0.1, 0.1]	$(x, \dot{x})$ $(\theta, \dot{\theta})$	$(x, \dot{x})$ $(z, \dot{z})$ $(\theta, \dot{\theta})$
Dynamics	white noise	std:[-0.1, 0.1]	pole length pole mass	quadrotor mass quadrotor inertia
Action	Impulse noise	Force: [-1, 1] Step offset: 20 Duration: 80 Decary rate: 0.9	horizontal forces	motors thrusts

C.3 More experiment results

C.3.1 Comparative Analysis of Fuzzy Operator and Min-Max Operator in Safe Reinforcement Learning

We first formally define three safety sets: the fundamental safety set $\mathcal{S}_c$ represents permissible state constraints, the safe forward invariant set $\mathcal{S}_\mathcal{I}$ (a subset of $\mathcal{S}_c$) guarantees persistent state containment within $\mathcal{S}_c$ under nominal conditions, and the robust safe forward invariant set $\mathcal{S}_\mathcal{R}$ (a conservative subset of $\mathcal{S}_\mathcal{I}$) maintains state invariance under worst-case disturbances.

As shown in Fig. 5(a), Conventional min-max approaches through robust control barrier functions (RCBFs) strictly confine states within $\mathcal{S}_\mathcal{R}$, where $\mathcal{S}_\mathcal{R}$ (yellow region) occupies only 23.6% of $\mathcal{S}_c$

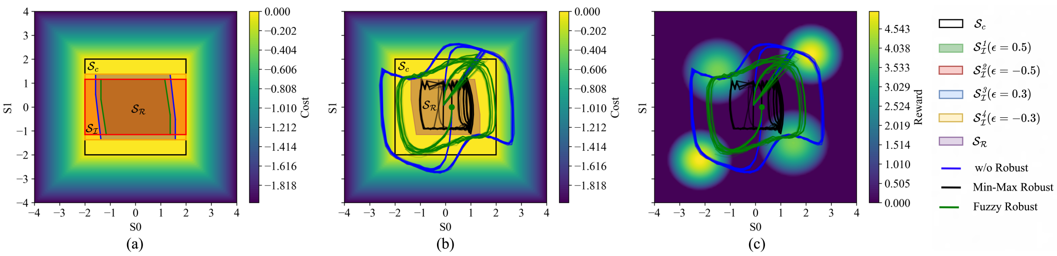

Figure 5: (a) Hierarchical relationship of safety sets, (b) Cost space and (c) Reward space trajectory comparisons, with dashed lines indicating safety boundaries.

(gray region). This conservative strategy ensures absolute safety at the cost of exploration capability, sacrificing access to 41.7% of high-reward regions.

Our proposed fuzzy robust method overcomes this limitation through dynamic weighting on different uncertainty levels. The training curves in Fig. 6 demonstrate that in the double-integrator environment, Fuz-RL's value iteration algorithm achieves 2.17× higher final returns compared to the min-max approach under Level-15 configuration. The underlying mechanism enables adaptive safety margin adjustment, permitting safe exploration in $\mathcal{S}_c \setminus \mathcal{S}_{\mathcal{R}}$ regions during 97.4% of test episodes. Trajectory heatmaps in Fig. 5(b)-(c) reveal that while conventional methods (black trajectories) remain strictly confined within $\mathcal{S}_{\mathcal{R}}$, and non-robust approaches (blue trajectories) risk 32.6% boundary violations, our fuzzy robust method (green trajectories) achieves optimal performance balance with 97% safety rate through dynamic fuzzy measure.

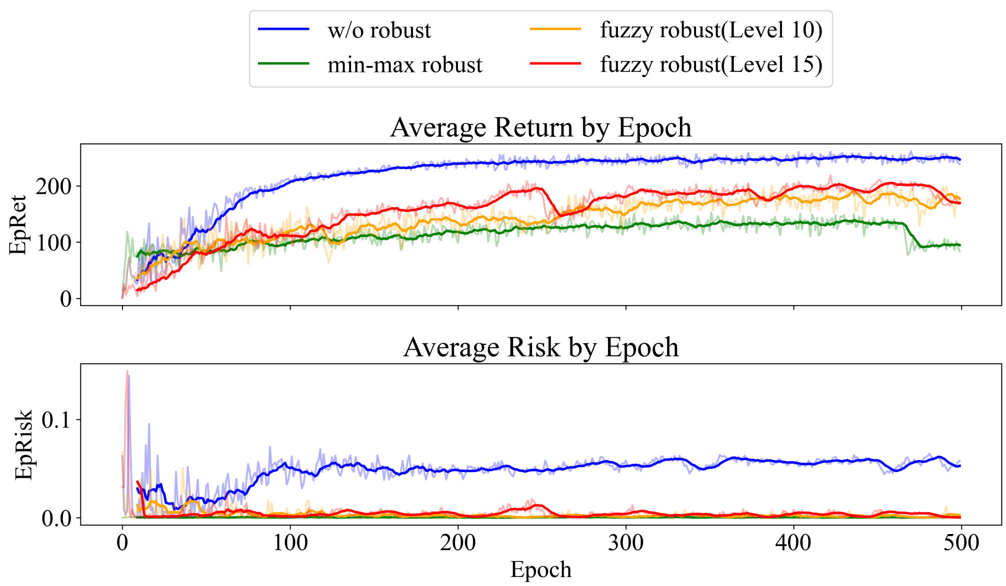

Figure 6: Training curve comparison in double integrator environment, with shaded regions indicating standard deviation across 5 random seeds.

C.3.2 Comparison between Safe RL and Fuz-RL

Similar to the Quadrotor-Track task, we set different levels of perturbations in the observation, dynamics, and action to evaluate the performance of the CartPole-Stab, CartPole-Track, and Quadrotor-Stab tasks under the three benchmark safe RL algorithms and the corresponding Fuz-RL, as shown in Figures 7, 8, and 9. Each point in the figures represents the average metrics from 10 episodes run for each of 10 different seeds.

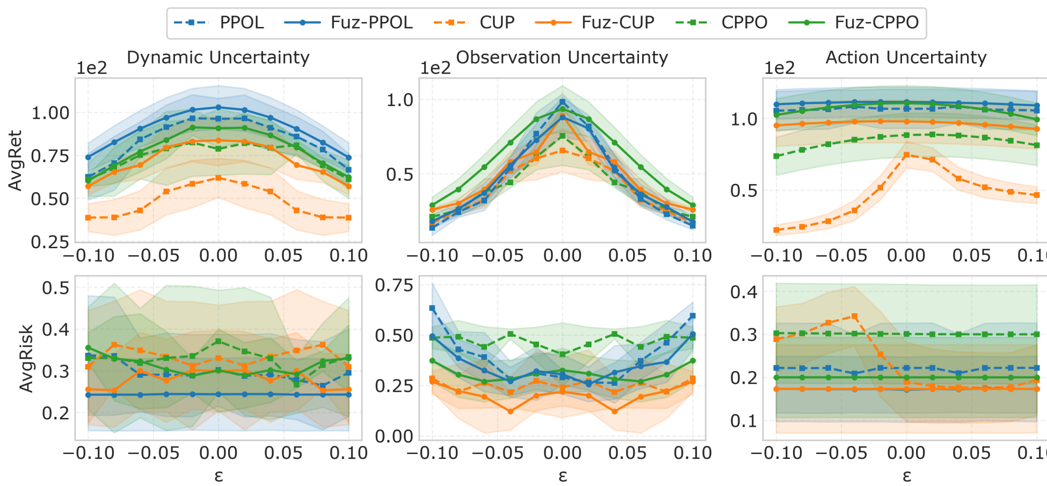

Figure 7: Average episodic rewards and average episodic risk of three safe RL and Fuz-RL under various uncertainty settings in Cartpole Stab task.

C.3.3 Comparison between Fuz-RL and RAMU

To compare with the current SOTA algorithms in robust safe RL, this section showcases the performance comparison between RAMU and Fuz-RL under uncertainties in observation, action, and dynamics, as depicted in Figures 11, Figures 12, and Figures 13, respectively.

C.3.4 Validation on Power System Frequency Control Task

The IEEE 39-bus system, a standard power grid benchmark with 10 generators and 46 transmission lines, was used to validate Fuz-RL's performance in frequency control tasks. The system state captures frequency deviations (Δf), generator rotor angles (δ), mechanical power outputs (P_m), and tie-line power flows (P_{tie}). Control actions involve real-time adjustments of generator active power setpoints (P_{ref}) and discrete load shedding commands (0-100% reduction). The primary objectives are to maintain frequency within $[59.8\,\text{Hz}, 60.2\,\text{Hz}]$ under stochastic load/renewable fluctuations while minimizing control costs ($\sum \|P_{\text{ref}} - P_{\text{nominal}}\|_2$) and avoiding safety-critical violations such as line overloads (>120% capacity).

Robustness tests were conducted under three uncertainty scenarios:

- *Observation noise* ($\sigma = 0.1\,\text{Hz}$ Gaussian noise in frequency measurements),
- *Action noise* (100ms delay + 5% bias in control signals),
- *Dynamics noise* ($\pm 10\%$ parameter drift in generator inertia/damping).

Table 4: Performance on IEEE 39-Bus Frequency Control (AvgRet / AvgRisk)

Case	Method	Observation Noise	Action Noise	Dynamic Noise
IEEE-39 Bus	PPOL	-5456.30 / 0.17	-6357.81 / 0.16	-7471.96 / 0.52
IEEE-39 Bus	**Fuz-PPOL**	**-4822.03 / 0.14**	**-5789.19 / 0.13**	**-7363.20 / 0.47**

Fuz-PPOL demonstrates consistent improvements over PPOL. Under observation noise, Fuz-PPOL get 11.6% higher returns and 17.6% lower risk. For action noise, the AvgRet metic is improved by 8.9% with 18.8% risk reduction. Under dynamics perturbations, Fuz-PPOL narrows performance degradation while reducing safety violations by 9.6%.

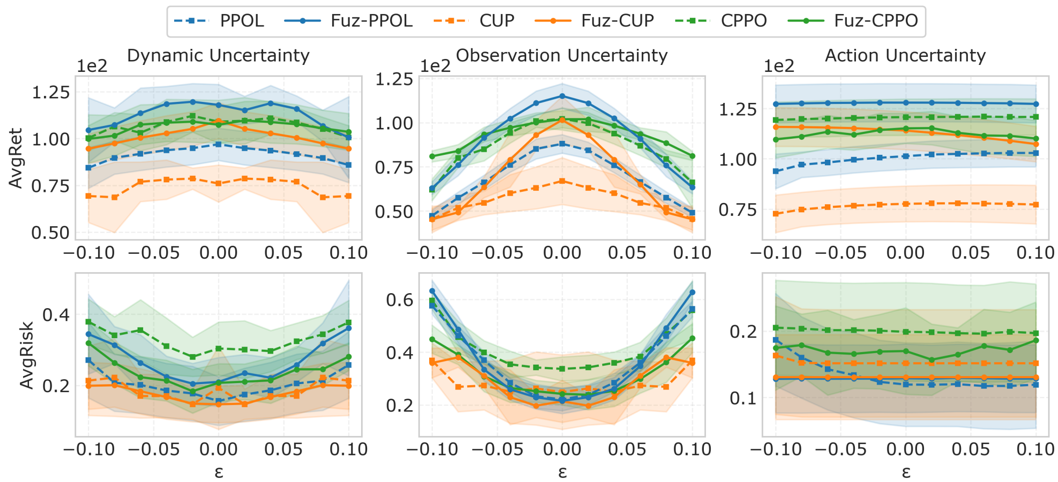

Figure 8: Average episodic rewards and average episodic risk of three safe RL and Fuz-RL under various uncertainty settings in Cartpole Track task.

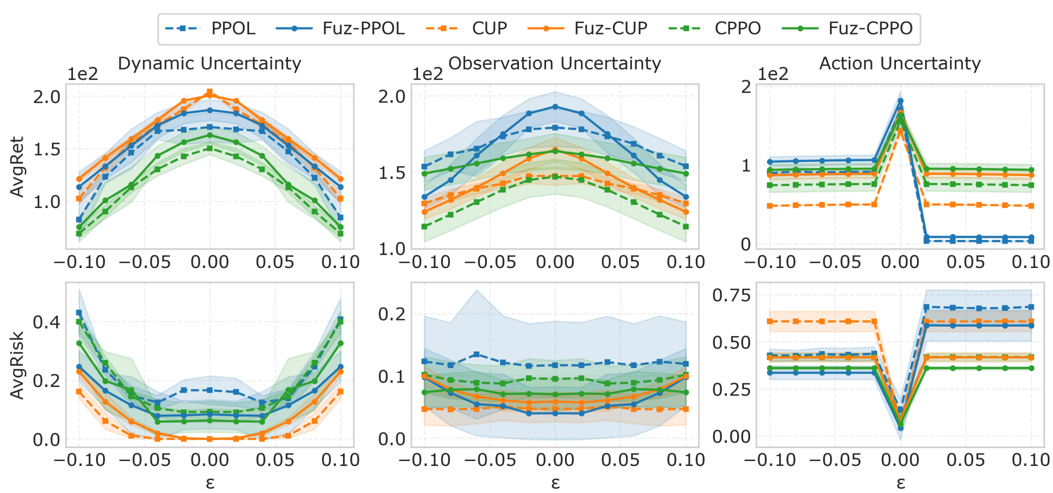

Figure 9: Average episodic rewards and average episodic risk of three safe RL and Fuz-RL under various uncertainty settings in Quadrotor Stab task.

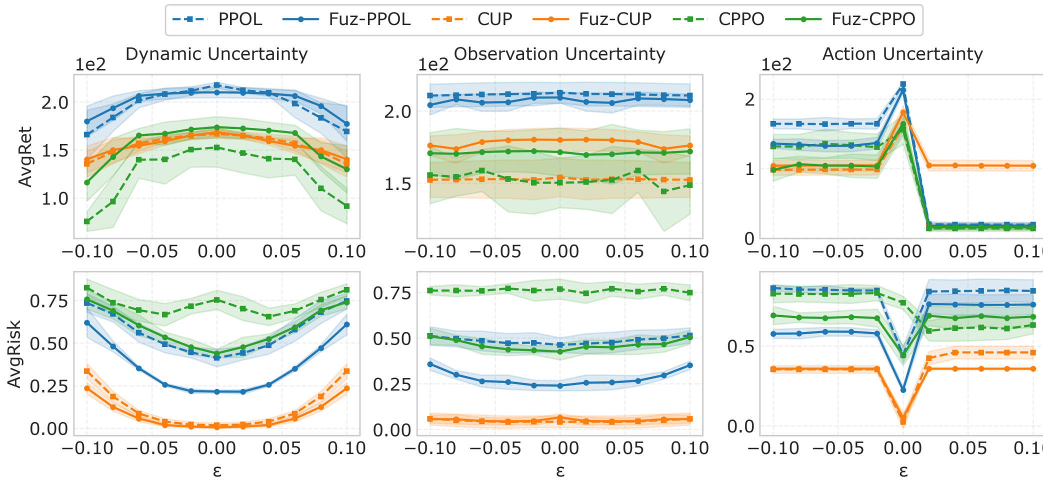

Figure 10: Average episodic rewards and average episodic risk of three safe RL and Fuz-RL under various uncertainty settings in Quadrotor Track task.

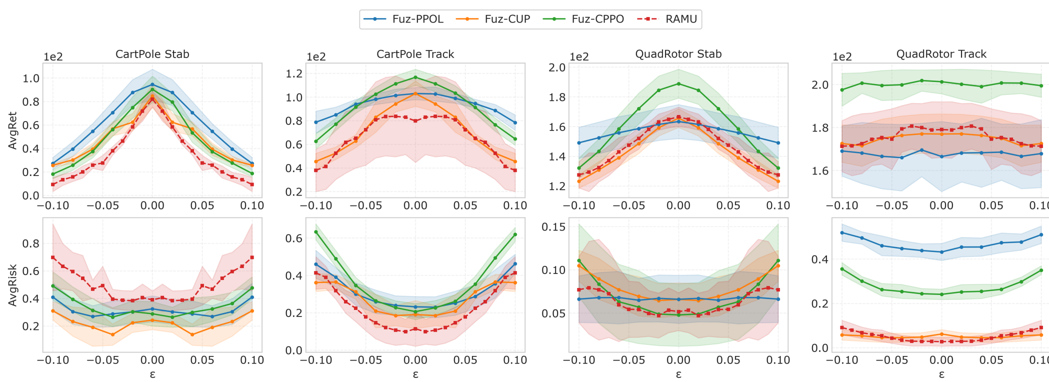

Figure 11: Average episodic reward (top) and average episodic risk (bottom) of Fuz-RL and RAMU in different scales' observation uncertainty settings. The horizontal axis represents the uncertainty level.

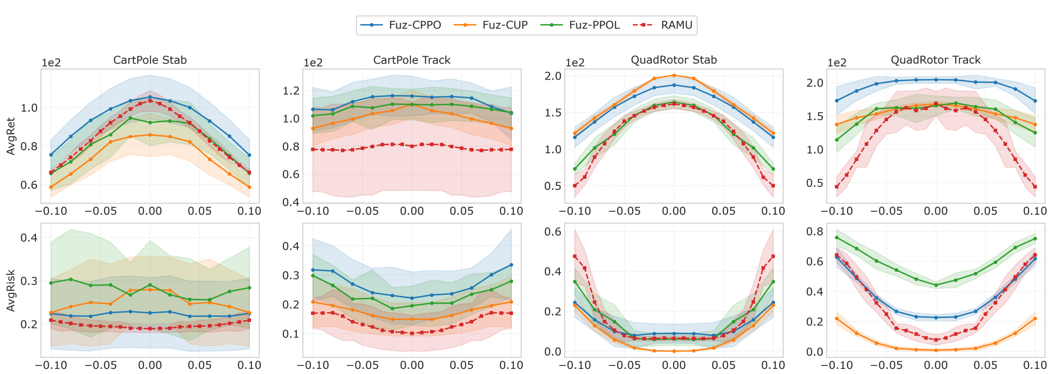

Figure 12: Average episodic reward (top) and average episodic risk (bottom) of Fuz-RL and RAMU in different scales' action uncertainty settings. The horizontal axis represents the uncertainty level.

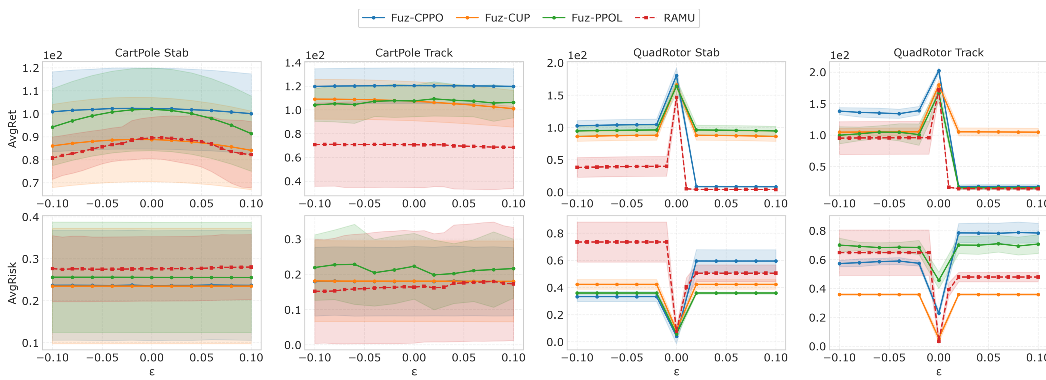

Figure 13: Average episodic reward (top) and average episodic risk (bottom) of Fuz-RL and RAMU in different scales' dynamics uncertainty settings. The horizontal axis represents the uncertainty level.

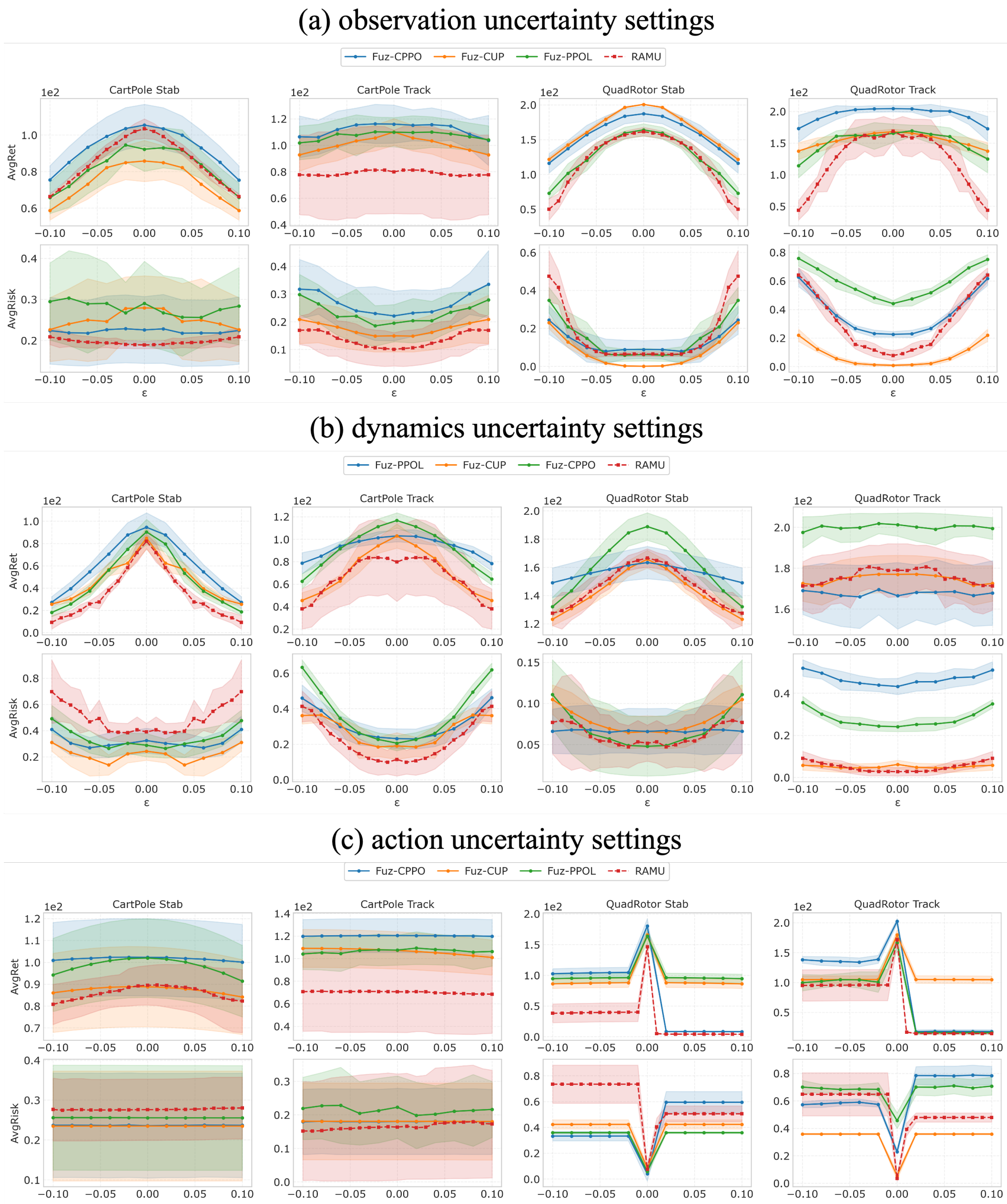

Figure 14: Average episodic reward (top) and average episodic risk (bottom) of Fuz-RL and RAMU in different scales' observation, dynamics, and action uncertainty settings. The horizontal axis represents the uncertainty level.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly indicate the main contributions of our Fuz-RL scope in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 6, we describe the limitations of the proposed Fuz-RL.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions of the theoretical result are clearly stated in the Theorem. 4.3, Theorem. 4.2 and Theorm. 4.4. Refer to Appendix A for the complete proof.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a thorough explanation in Section C.1 of all our experiments' settings, and the more general setting and detailed results can be found in Appendix C.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The whole code is provided in the supplemental material with sufficient instructions in the "README.md" file.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix C.2 clearly states the implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We evaluate our experimental results across 10 episodes with 10 different random seeds, reporting standard deviations as error bars in the figures and as variance metrics in the tables.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We indicate the information of compute resources in appendix C.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We strictly adhere to the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The potential impacts are stated in the Conclusion, see Section 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All codes of Fuz-RL are implemented based on the SpinningUp [1] which is explicitly mentioned and properly respected in Section 5.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets are attached in the supplemental material.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not include crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve any crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in Fuz-RL does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.