Uncertainty-Aware Safety-Critical Decision and Control for Autonomous Vehicles at Unsignalized Intersections

Ran Yu, Zhuoren Li, Lu Xiong, Wei Han, Bo Leng

***Abstract*— Reinforcement learning (RL) has demonstrated potential in autonomous driving (AD) decision tasks. However, applying RL to urban AD, particularly in intersection scenarios, still faces significant challenges. The lack of safety constraints makes RL vulnerable to risks. Additionally, cognitive limitations and environmental randomness can lead to unreliable decisions in safety-critical scenarios. Therefore, it is essential to quantify confidence in RL decisions to improve safety. This paper proposes an Uncertainty-aware Safety-Critical Decision and Control (USDC) framework, which generates a risk-averse policy by constructing a risk-aware ensemble distributional RL, while estimating uncertainty to quantify the policy's reliability. Subsequently, a high-order control barrier function (HOCBF) is employed as a safety filter to minimize intervention policy while dynamically enhancing constraints based on uncertainty. The ensemble critics evaluate both HOCBF and RL policies, embedding uncertainty to achieve dynamic switching between safe and flexible strategies, thereby balancing safety and efficiency. Simulation tests on unsignalized intersections in multiple tasks indicate that USDC can improve safety while maintaining traffic efficiency compared to baselines.**

I. Introduction

Autonomous driving (AD) is becoming a significant focus of global innovation due to its potential in energy saving, emission reduction, traffic efficiency, road safety, and releasing drivers [1]. Although rule-based autonomous vehicles (AV) are effective in reducing accident rates in straightforward, merge, and exit scenarios, the accident frequency in turning scenarios is 1.98 times higher than that of human-driven vehicles (HDVs) [2]. Due to the limitations of predefined rules in covering all driving scenarios, AVs struggle to make flexible decisions when faced with multiple oncoming HDVs, potentially causing traffic congestion or rear-end collisions [3], [4].

Recently, reinforcement learning (RL) has demonstrated remarkable proficiency in decision-making tasks, such as those on the highway [5], merge [6], and intersection [7], due to its ability to optimize high-dimensional state spaces through continuous interaction with dynamic environments. However, even well-trained RL agents encounter considerable challenges in ensuring the safety of policies due to the lack of safety constraints [8]. Consequently, Safe RL has emerged as a paradigm for RL applications, with the objective of maximizing the expected cumulative reward while simultaneously satisfying safety constraints [9].

A typical Safe RL architecture is based on the Constrianed Markov Decision Process (CMDP) framework, which guarantees safety by constraining the cumulative expected cost below a threshold. This is usually addressed by solving the constrained optimization problem using Lagrangian multiplier methods [10], [11] or trust region methods [12], [13]. However, as the policy gradually converges near the safety threshold, learning a safe policy becomes more challenging due to the sparsity of the cost signal. Another common paradigm for Safe RL involves performing safety corrections to the RL policy, such as using action masks [14] and safe energy functions [15]–[17]. Action masks identify unsafe actions by predefining a safe action set and limiting unsafe actions from participating in the network's updates. Nevertheless, such methods are typically applicable only to discrete action spaces. Control barrier functions (CBF) are a significant representative of safety energy functions and are widely regarded as a well-established method in safety control [18]. However, many CBF approaches require the design of intricate prior functions, which limits their adaptability and generalization in complex dynamic scenarios. Furthermore, the existence of high relative degree between the standard CBF and the underlying system control inputs requires a more conservative barrier function for safety, otherwise the safety can be compromised due to the inability to constrain the higher-order dynamics resulting in transient transgressions of the lower-order constraints. Learning-based safe energy functions have garnered significant attention because of their independence from prior knowledge. However, they are deficient in clearly defined and interpretable safety guarantee mechanisms.

RL decisions typically yield only outcomes without revealing the uncertainty. This inherent ambiguity can pose safety risks in AD tasks [19]. This is particularly evident in scenarios involving long-tailed problems or out-of-distribution (OOD) events, where the reliability of RL policies is significantly compromised. The quantification of the confidence level of RL decisions facilitates the identification of RL actions that pose potential safety risks due to high uncertainty. This, in turn, enables the reduction or avoidance of dangerous decisions in OOD scenarios. Uncertainty can be categorized into two main types: epistemic uncertainty (EU) and aleatoric uncertainty (AU) [20]. EU reflects data-dependent model or parameter uncertainty, which arises from lack of knowledge and can be reduced by observing more data; AU is irreducible, reflecting the stochastic nature of

Ran Yu, Zhuoren Li, Lu Xiong, Wei Han and Bo Leng are with the School of Automotive Studies, Tongji University, Shanghai 201804, China (email:ranyu@tongji.edu.cn; 1911055@tongji.edu.cn; xiong_lu@tongji.edu.cn; tjhanwei@tongji.edu.cn; lengbo@tongji.edu.cn;).

environmental dynamics [21]. Recent works in the field of AD have demonstrated that uncertainty-aware RL has garnered significant attention and focus. A common approach is to use a backup policy to replace the RL policy when the estimated uncertainty exceeds a threshold or falls outside a safety set [8], [22]–[25]. However, the set or threshold is often limited to specific scenarios. Furthermore, some backup strategies, such as deceleration or braking, may negatively impact traffic efficiency.

To this end, in order to enhance the ability of RL to handle safety-critical and OOD scenarios while achieving a balance between safety and traffic efficiency, we propose the Uncertainty-aware Safety-critical Decision and Control (USDC) framework, as illustrated in Fig. 1. This framework leverages deep ensemble (DE) [26] to estimate the joint uncertainty (JU) that encompasses both EU and AU. By dynamically adjusting safety constraints based on JU, it effectively balances safety and efficiency. The main contributions are as follows:

- We propose a risk-aware distributional RL ensemble architecture that combines the original and Fixed Prior Networks (FPN) to build the critic. It quantifies tail risk in the reward distribution and generates a risk-averse policy, while jointly estimating EU and AU for dual uncertainty-driven decision-making.
- A high-order control barrier function (HOCBF) is constructed, which ensures safety using only relative distance. Moreover, HOCBF incorporates JU to dynamically adjust safety constraints, balancing safety and traffic efficiency.
- Based on the JU distribution, the HOCBF and RL policies are evaluated, ensuring that the RL policy performs no worse than the HOCBF under low uncertainty and favors safer policies under high uncertainty.

II. Preliminaries

A. Distributional Reinforcement Learning

RL can be modeled as a Markov Decision Process (MDP) defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \rho_0, \gamma)$. $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition probability function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\rho_0 : \mathcal{S} \to [0,1]$ is the initial state distribution, and γ is the discount factor for future reward. The policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ is a map from given states to a probability distribution on action space. Standard MDP aims at maximizing the agent's cumulative discounted reward $\mathcal{L}(\pi) = \mathbb{E}_{\pi,\mathcal{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$. For $\pi \in \Pi$, the Q function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ can be defined as $Q^\pi(s,a) := \mathbb{E}_{\pi,\mathcal{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$, which satisfies the following Bellman equation:

$$Q(s,a) := \mathbb{E}[r(s,a)] + \gamma \mathbb{E}_{\pi,\mathcal{P}}[Q(s',a')]. \tag{1}$$

In the distributional RL setup, the distribution over returns $Z(s,a)$ is estimated, and its expectation is maximized to obtain $Q(s,a) = \mathbb{E}[Z(s,a)]$. Hence, we can rewrite (1) as:

$$Z(s,a) \overset{D}{:=} r(s,a) + \gamma Z(s',a'), \tag{2}$$

where $\overset{D}{:=}$ indicates that both sides of the equation are distributed according to the same distribution. Let $F_z(z) = \mathbb{P}(Z \le z)$ denote the cumulative distribution function (CDF) of random variable Z, the quantile function F_Z^{-1} can be represented as the inverse of CDF. Given quantile fraction τ, we have $F_Z^{-1} := \inf\{z \in \mathbb{R} : \tau \le F_Z(z)\}$.

Following [27], we defined a serious of quantile fractions $\{\tau_i\}_{i=0,\dots,N}$, $\hat{\tau}_i = (\tau_i + \tau_{i+1}/2)$, and the pairwise temporal difference (TD) error between two quantile fractions $\hat{\tau}_i, \hat{\tau}_j$:

$$\delta_{ij} = r + \gamma \left[Z^{\bar{\theta}}_{\hat{\tau}_i}(s',a') - \alpha \log \pi_{\bar{\phi}}(a'|s') \right] - Z^{\theta}_{\hat{\tau}_j}(s,a) \tag{3}$$

where $\bar{\theta}, \theta, \bar{\phi}$ are parameters of target critic, critic, and target actor network, respectively. α is the temperature parameter. We then tarin the critic $Z^\theta_\tau(s,a)$ using quantile regression by minimizing the weighted pairwise Huber loss [28] with threshold κ:

$$\mathcal{L}_Z(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s') \in \mathcal{B}} \sum_{i=0}^{N_q-1} \sum_{j=0}^{N_q-1} (\tau_{i+1} - \tau_i) \rho^\kappa_{\hat{\tau}_j}(\delta_{ij}) \tag{4}$$

$$\rho^\kappa_\tau(\delta_{ij}) = |\tau - \mathbb{I}\{\delta_{ij} < 0\}| \frac{\mathcal{L}_\kappa(\delta_{ij})}{\kappa}, \tag{5}$$

$$\mathcal{L}_\kappa(\delta_{ij}) = \begin{cases} \frac{1}{2}\delta_{ij}^2, & \text{if } |\delta_{ij}| \le \kappa \\ \kappa\left(|\delta_{ij}| - \frac{1}{2}\kappa\right), & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathcal{B}$ is a minibatch of transitions sampled from a replay buffer, N_q denotes the number of quantile points over $[0,1]$. Then, the objective of actor is to maximize the Q-return:

$$L_\pi(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}, a_t \sim \pi_\phi}[\alpha \log(\pi_\phi(a_t|s_t)) - Q_\theta(s_t,a_t)], \tag{7}$$

where $Q_\theta(s,a) = \sum_{i=0}^{N_q} (\tau_{i+1} - \tau_i) Z^\theta_{\hat{\tau}_i}(s,a)$, $\mathcal{D}$ is the transitions replay buffer.

B. High-Order Control Barrier Functions

Consider an input-affine control system given by:

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}) + g(\boldsymbol{x})\boldsymbol{u}, \tag{8}$$

where the system state $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^n$ and the control input $\boldsymbol{u} \in \mathcal{U} \subset \mathbb{R}^{u_n}$. f and g are locally Lipschitz.

Definition 1 (Forward invariant set): The set $\mathcal{C}$ is forward invariant for system (8) if for every initial condition $\boldsymbol{x}(t_0) \in \mathcal{C}$, $\boldsymbol{x}(t) \in \mathcal{C}$ for $\forall t \ge t_0$. For a continuously differentiable function $h : \mathcal{X} \to \mathbb{R}$, let

$$\mathcal{C} = \{\boldsymbol{x} \in \mathcal{X} : h(\boldsymbol{x}) \ge 0\}. \tag{9}$$

Definition 2 (CBF [18]): Denote class $\mathcal{K}$ function is a function $\alpha : [0,a) \to [0,\infty), a > 0$ that is strictly increasing and with $\alpha(0) = 0$. Given the superlevel set $\mathcal{C}$ as in (9), h is a CBF if there exists a class $\mathcal{K}$ function α such that

$$\sup_{\boldsymbol{u} \in \mathcal{U}} [L_f h(\boldsymbol{x}) + L_g h(\boldsymbol{x})\boldsymbol{u}] \ge -\alpha(h(\boldsymbol{x})), \forall \boldsymbol{x} \in \mathcal{C}, \tag{10}$$

where $L_f h, L_g h$ denote the Lie derivatives along f and g, respectively. In the discrete system perspective, we consider a discrete-time system $\boldsymbol{x}_{k+1} = f(\boldsymbol{x}_k) + g(\boldsymbol{x}_k)\boldsymbol{u}_k$, where $k \in \mathbb{N}$ denotes the time step. Then, given a set $\mathcal{C}$ as in (9), continuous function h is a candidate discrete CBF if there exists a class $\mathcal{K}$ function satisfying $\alpha(x) \le x$ such that:

$$\sup_{\boldsymbol{u} \in \mathcal{U}} [\Delta h(\boldsymbol{x}, \boldsymbol{u}) + \alpha(h(\boldsymbol{x}))] \ge 0, \forall \boldsymbol{x} \in \mathcal{C}, \tag{11}$$

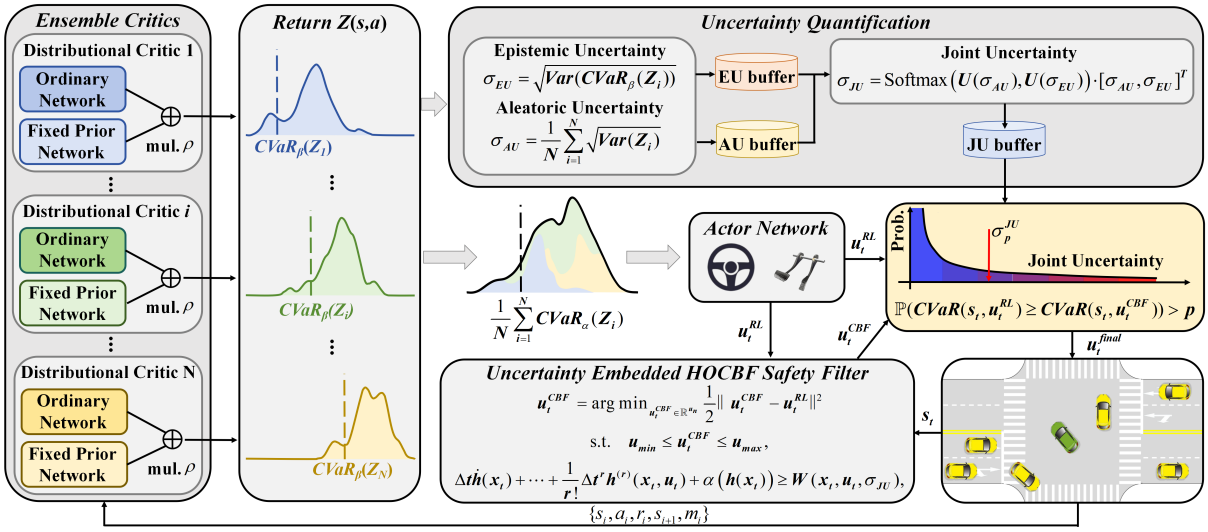

Fig. 1. Diagram of our framework.

where $\Delta h(\boldsymbol{x}_k, \boldsymbol{u}_k) = h(\boldsymbol{x}_{k+1}) - h(\boldsymbol{x}_k)$. Note that linear class $\mathcal{K}$ function is commonly used in the discrete domain, i.e., $\alpha(h(\boldsymbol{x})) = \lambda h(\boldsymbol{x}), \lambda \in (0, 1]$.

Definition 3 (Relative degree): A continuously differentiable function $h : \mathcal{X} \to \mathbb{R}$ is said to have relative degree $r \in \mathbb{N}$ with respect to system (1), if $\forall x \in \mathcal{X}, L_g L_f^i h(\boldsymbol{x}) = 0, \forall i \in [0, 1, ..., r-2]$, and $L_g L_f^{r-1} h(\boldsymbol{x}) \neq 0$.

The standard CBF assumes a relative degree of 1, meaning that the first derivative of the safety constraint is equal to the control input. However, this doesn't apply to many autonomous driving constraints, such as the relative degree between distance and acceleration is 2. Therefore, high-order CBFs (HOCBF) are needed to handle such problems.

For the system and the original function h with relative degree r, we define the auxiliary function $\Psi_0(x) := h(\boldsymbol{x})$ and derive the following recursively:

$$\Psi_i(x) := \dot{\Psi}_{i-1}(x) + \alpha_i(\Psi_{i-1}(x)), \forall i \in \{1, \dots, r\}, \quad (12)$$

where each α_i is a class $\mathcal{K}$ function. Then, a series of superlevel set $\mathcal{C}_i$ can be represented as:

$$\mathcal{C}_i := \{x \in \mathcal{X} : \Psi_{i-1}(x) \geq 0\}, \quad \forall i \in \{1, \dots, r\}. \quad (13)$$

Definition 4 (HOCBF [29], [30]): Given a serious sets $\mathcal{C}_i$ as in (13), a continuously differentiable function $h : \mathcal{X} \to \mathbb{R}$ is a candidate HOCBF with relative degree r if there exist class $\mathcal{K}$ functions $\alpha_i, \forall i \in \{1, \dots, r\}$ such that

$$\sup_{\boldsymbol{u} \in \mathcal{U}} \Psi_r(\boldsymbol{x}, \boldsymbol{u}) \geq 0, \quad \forall \boldsymbol{x} \bigcap_{i=1}^{r} \mathcal{C}_i \quad (14)$$

III. Methodologies

A. Uncertainty-aware Risk-sensitive Distributional RL

1) Risk sensitive RL: In safety-critical scenarios, we want the AV to adopt conservative measures in the face of high uncertainty, as increased uncertainty typically leads to higher risks. Specifically, when binary safety indicators, such as collision occurrence, are involved, the AV's estimated return distribution tends to follow a multimodal distribution: there are both high return modes corresponding to desirable behaviors and low return modes for unsafe behaviors. To enhance safety, the objective is to minimize the occurrence of low return modes. Thus, we incorporate conditional value-at-risk (CVaR) [31] to formulate a risk-sensitive policy:

$$CVaR_\beta(Z) = \mathbb{E}\left[Z | Z < VaR_\beta(Z)\right], \quad (15)$$

where $VaR_\beta(Z)$ (the value at risk) is the β-worst percentile of Z. Denote distortion function $\zeta : [0, 1] \to [0, 1]$, which is strictly increasing and satisfies $\zeta(0) = 0$ and $\zeta(1) = 1$. The distorted expectation of Z can be expressed as $\Xi(Z) = \int_0^1 F_Z^{-1}(\tau) d\zeta(\tau) = \int_0^1 \zeta'(\tau) F_Z^{-1}(\tau) d\tau$. And the $CVaR$ is given by $CVaR_\beta(Z) = \sum_{i=0}^{N_q - 1} (\tau_{i+1} - \tau_i) \zeta'(\hat{\tau}_i) Z^\theta_{\hat{\tau}_i}(s, a)$, where $\zeta(\tau) = \min\{\tau/\beta, 1\}$.

2) Uncertainty Quantification: We introduce deep ensembles [26] to estimate uncertainty. To improve the recognition ability for OOD scenarios, it is anticipated that the ensemble members demonstrate consistent performance on in-distribution data, whilst preserving diversity on OOD data [32]. Specifically, we construct an ensemble critic architecture consisting of N groups of critics Z_n^θ and target critics $Z_n^{\bar{\theta}}$, $n \in [1, .., N]$. Bootstrapping [33] was utilized to ensure that each ensemble member has access to a unique subset of the experience replay buffer. Furthermore, we introduce a Randomized Prior Function (RPF) [34], which adds a fixed prior network (FPN) of the same size as the original network to each ensemble member, improving Bayesian posterior estimates and enhancing diversity on OOD data. The Z-return of each ensemble member is composed of two parts:

$$Z^\theta_{n,\tau}(s, a) = \frac{\mathcal{O}(s, a; \theta_n) + \rho \mathcal{F}(s, a; \tilde{\theta}_n)}{1 + \rho}, \quad (16)$$

where $\mathcal{O}(\cdot), \mathcal{F}(\cdot)$ stand for original network and FPN, respectively. $\tilde{\theta}_n$ are fixed parameters of FPN, ρ is prior factor.

Then, the TD-error of n-th critic is:

$$\delta_{ij}^{n} = r + \gamma \left[\bar{Z}_{\widehat{\tau}_i}^{\bar{\theta}}(s', a') - \alpha \log \pi_{\bar{\phi}}(a'|s') \right] - Z_{n,\widehat{\tau}_j}^{\theta}(s, a), \tag{17}$$

where $\bar{Z}_{\widehat{\tau}_i}^{\bar{\theta}}(s', a') = \frac{1}{N} \sum_{n=1}^{N} \bar{Z}_{n,\widehat{\tau}_i}^{\bar{\theta}}(s', a')$ is the average Z-return of N target critic, $a' \sim \pi_{\bar{\phi}}(s')$. The quantile Huber regression loss in (4) is then rewritten as:

$$\mathcal{L}_Z^n(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s') \in \mathcal{B}} \sum_{i=0}^{N_q-1} \sum_{j=0}^{N_q-1} m_n (\tau_{i+1} - \tau_i) \rho_{\widehat{\tau}_j}^{\kappa} \left(\delta_{ij}^{n} \right), \tag{18}$$

where $m_n \sim Bernoulli(\mathrm{p}), \mathrm{p} \in (0, 1], n \in [1, ..N]$ is bootstrap masks sample from Bernoulli distribution. Similarly, the actor objective can be modified as:

$$L_\pi(\phi) = \mathbb{E}_{s_t \sim \mathcal{D}, a_t \sim \pi_\phi} [\alpha \log(\pi_\phi(a_t|s_t)) - \overline{CVaR}_\beta(s, a)], \tag{19}$$

where $\overline{CVaR}_\beta(s, a) = \frac{1}{N} \sum_{n=1}^{N} CVaR_\beta(Z_n^\theta)$.

In the context of this paper, epistemic uncertainty is estimated by the disagreement between individual members and the overall ensemble, and is represented through the standard deviation of $CVaR$:

$$\sigma_{EU} = \sqrt{Var(CVaR_\beta(Z_n^\theta))}. \tag{20}$$

Similar to the approach in [22], we represent aleatoric uncertainty by estimating the standard deviation of the Z-return of the ensemble critic:

$$\sigma_{AU} = \frac{1}{N} \sum_{n=1}^{N} \sqrt{Var(Z_n^\theta)}. \tag{21}$$

In accordance with the estimated AU and EU, the utilization of joint uncertainty (JU) is employed to account for their combined effect:

$$\sigma_{JU} = \mathrm{Softmax}\left(U(\sigma_{AU}), U(\sigma_{EU})\right) \cdot [\sigma_{AU}, \sigma_{EU}]^\top, \tag{22}$$

where $U(x) = 1/2(\tanh(x - \mu)/\eta + 1)$, μ, η are the mean and variance of σ_{EU} or σ_{AU}, calculated from their buffer, which collects $\mathcal{M}$ steps of σ_{EU} and σ_{AU}. Then, σ_{JU} will also be added to the corresponding buffer.

B. Uncertainty-embedded HOCBF

1) Safety Filter Formulation: After obtaining the risk-averse policy $\boldsymbol{u}_k^{RL}$, we introduce HOCBF as a safety filter to further enhance safety. Kinematic bicycle model [35] is utilized to describe the motion of AV:

$$\boldsymbol{x}_{k+1} = \begin{bmatrix} x_k + \Delta t \cdot v_k \cos(\varphi_k) \\ y_k + \Delta t \cdot v_k \sin(\varphi_k) \\ v_k + \Delta t \cdot a_{\text{lon}} \\ \varphi_k + \Delta t \cdot v_k \tan \delta_f / L \end{bmatrix} \tag{23}$$

where $\boldsymbol{x} = [x, y, v, \varphi]^\top$, x, y are the position coordinates, $v, \varphi, a_{\text{lon}}, \delta_f, \Delta t, L$ are velocity, heading angle, longitudinal acceleration, steering angle of front wheel, sample time and wheelbase, respectively. Based on the relative distance depicted in Fig. 2, we select the candidate CBF as follows:

$$\begin{aligned} h_{\text{veh}}(\boldsymbol{x}_k) &= (x_k - x_k^{\text{obs}})^2 + (y_k - y_k^{\text{obs}})^2 - (r_{obs} + r_{ego})^2, \\ h_{\text{road}}(\boldsymbol{x}_k) &= (x_k - x_k^{\text{road}})^2 + (y_k - y_k^{\text{road}})^2 - r_{ego}^2, \end{aligned} \tag{24}$$

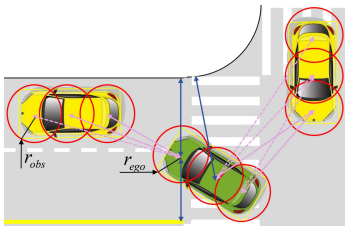

Fig. 2. Diagram of relative distance. At the k-th time step, the envelopes of the ego vehicle (EV) and surrounding vehicles (SV) are represented by three circles, denoted as $(x_{i|k}^{\text{ego}}, y_{i|k}^{\text{ego}})$ and $(x_{i|k}^{\text{obs}}, y_{i|k}^{\text{obs}})$, $i \in \{1, 2, 3\}$. Then, we select $\mathbf{N}$ closest points to the EV, $(x_{j|k}^{\text{obs}}, y_{j|k}^{\text{obs}})$, $j \in \{1, ..., \mathbf{N}\}$, and $\mathbf{M}$ constraints related to the road boundaries. The construction of distance constraints are formulated as a total of $\{\mathbf{N} \times 3 + \mathbf{M}\}$ constraints.

where $(x_k^{\text{road}}, y_k^{\text{road}})$ denote the position of road boundary point, and the candidate CBF has relative degree $r = 2$.

Consider a discrete-time system as described in (23). To address the problem of high relative degree, we introduce the Truncated Taylor CBF (TTCBF) [36], which approximates the discrete-time CBF condition using a truncated Taylor series and requires only a class $\mathcal{K}$ function. For a system with relative degree r, TTCBF approximates $\Delta h(\boldsymbol{x}_k, \boldsymbol{u}_k)$ in (11) as $\Delta h(\boldsymbol{x}_k, \boldsymbol{u}_k) \approx \Delta t \dot{h}(\boldsymbol{x}_k) + \frac{1}{2}\Delta t^2 \ddot{h}(\boldsymbol{x}_k) + \cdots + \frac{1}{r!}\Delta t^r h^{(r)}(\boldsymbol{x}_k, \boldsymbol{u}_k)$, with the r-th derivative $h^{(r)}(\boldsymbol{x}_k, \boldsymbol{u}_k)$ capture the control input.

Definition 5 (TTCBF [36]): Given a set $\mathcal{C}$ as in (9), a continuously differentiable function h is a candidate TTCBF with relative degree r if there exist class $\mathcal{K}$ functions $\alpha(x) \leq x$ such that $\forall \boldsymbol{x}_k \in \mathcal{C}$,

$$\begin{aligned} \sup_{\boldsymbol{u}_k \in \mathcal{U}} [\Delta t \dot{h}(\boldsymbol{x}_k) + \cdots + \frac{1}{r!}\Delta t^r h^{(r)}(\boldsymbol{x}_k, \boldsymbol{u}_k) + \\ \alpha\left(h(\boldsymbol{x}_k)\right)] \geq \Gamma \Delta t^{r+1}, \end{aligned} \tag{25}$$

where Γ is a hyper-parameter satisfying $\frac{\Gamma}{(r+1)!} t^{r+1} \geq |R_{r+1}| = \frac{|h^{(r+1)}(\boldsymbol{\xi})|}{(r+1)!} t^{r+1}, \boldsymbol{\xi} \in [\boldsymbol{x}_k, \boldsymbol{x}_{k+1}]$.

The first and second time derivatives of h can be derived from (23) and (24) as:

$$\begin{aligned} \dot{h}(\boldsymbol{x}_k) &= 2(x_k - x_k^{\text{point}}) v \cos\varphi + 2(y_k - y_k^{\text{point}}) v \sin\varphi, \\ \ddot{h}(\boldsymbol{x}_k) &= 2v^2 + 2a_{\text{lon}}\Delta_{\text{lon}} + 2v^2 \tan\delta_f \Delta_{\text{lat}}/L, \end{aligned} \tag{26}$$

where point $\in$ [obs, road], $\Delta_{\text{lon}} = (x_k - x_k^{\text{point}}) \sin\varphi + (y_k - y_k^{\text{point}}) \cos\varphi$, $\Delta_{\text{lat}} = -(x_k - x_k^{\text{point}}) \sin\varphi + (y_k - y_k^{\text{point}}) \cos\varphi$.

Then, substitute $\dot{h}(\boldsymbol{x}_k), \ddot{h}(\boldsymbol{x}_k)$ into (25):

$$\begin{aligned} &\Delta t \dot{h}(\boldsymbol{x}_k) + \frac{1}{2!}\Delta t^2 \ddot{h}(\boldsymbol{x}_k, \boldsymbol{u}_k) + \alpha\left(h(\boldsymbol{x}_k)\right) \geq \Gamma \Delta t^3, \\ \Rightarrow\ &\Delta t^2 \left[a_{\text{lon}}\Delta_{\text{lon}}, v^2 \Delta_{\text{lat}}/L\right] \begin{bmatrix} a_{\text{lon}} \\ \tan\delta \end{bmatrix} + h_{\text{cst}} - \Gamma \Delta t^3 \geq 0, \end{aligned} \tag{27}$$

where $h_{\text{cst}} = \Delta t \dot{h}(\boldsymbol{x}_k) + \alpha\left(h(\boldsymbol{x}_k)\right) + \Delta t^2 v^2$.

The above inequality (27) is denoted by $\mathcal{H}$. Similarly, we derive the inequality for all $\{\mathbf{N} \times 3 + \mathbf{M}\}$ constraints. The objective of the safety filter is to ensure safety constraints while minimizing modifications to $\boldsymbol{u}_k^{RL}$. Consequently, the

following optimization problem can be formulated:

$$\begin{aligned}\boldsymbol{u}_k^{CBF} =& \arg\min_{\boldsymbol{u}\in\mathcal{U}} \quad \frac{1}{2}\|\boldsymbol{u}_k^{CBF} - \boldsymbol{u}_k^{RL}\|^2 \\ &\text{s.t.} \quad \boldsymbol{u}_{\min} \le \boldsymbol{u}_k^{CBF} \le \boldsymbol{u}_{\max}, \\ \mathcal{H}_{\text{veh}} =& [\mathcal{H}_{\text{veh}}^{1}, .., \mathcal{H}_{\text{veh}}^{\mathbf{N}}, \mathcal{H}_{\text{veh}}^{\mathbf{N+1}}, ..., \mathcal{H}_{\text{veh}}^{\mathbf{2N}}, \mathcal{H}_{\text{veh}}^{\mathbf{2N+1}}, ..., \mathcal{H}_{\text{road}}^{\mathbf{3N}}]^\top \\ \mathcal{H}_{\text{road}} =& [\mathcal{H}_{\text{road}}^{1}, .., \mathcal{H}_{\text{road}}^{\mathbf{M}}]^\top,\end{aligned} \tag{28}$$

where $\boldsymbol{u}_{\min}, \boldsymbol{u}_{\max}$ represent the input constraints. The optimization problem will be solved using the CasADi nonlinear programming solver [37].

2) Uncertainty-embedded Constraints and Probabilistic Improvement: We expect AV tend to execute safe maneuvers at higher uncertainty, while allowing for greater flexibility under low uncertainty, rather than always executing conservative maneuvers and thus sacrificing passage efficiency. Consider an uncertain system dynamics, $\dot{\boldsymbol{x}} = \hat{f}(\boldsymbol{x}) + \hat{g}(\boldsymbol{x})\boldsymbol{u} + \varphi(\boldsymbol{x}, \boldsymbol{u})$, where $\hat{f}, \hat{g}$ are known nominal dynamics, $\varphi(\boldsymbol{x}, \boldsymbol{u}) \in \mathcal{X} \times \mathcal{U} \subset \mathbb{R}^n$ is the uncertain vector to be estimated. The uncertainty is then embedded into CBF constraint, such that $\forall \boldsymbol{x} \in \mathcal{C}$:

$$\sup_{\boldsymbol{u}\in\mathcal{U}} \inf_{w\in\mathcal{W}} \left[L_{\hat{f}} h(\boldsymbol{x}) + L_{\hat{g}} h(\boldsymbol{x})\boldsymbol{u} + w(\boldsymbol{x}, \boldsymbol{u}) \right] \ge -\alpha(h(\boldsymbol{x})), \tag{29}$$

where $w(\boldsymbol{x}, \boldsymbol{u}) = L_\varphi h$ is the uncertain scalar. Then, (25) can be rewritten as:

$$\begin{aligned}\sup_{\boldsymbol{u}_k\in\mathcal{U}} \inf_{w\in\mathcal{W}} [\Delta t \dot{h}(\boldsymbol{x}_k) + \cdots + \frac{1}{r!}\Delta t^r h^{(r)}(\boldsymbol{x}_k, \boldsymbol{u}_k) + \\ w(\boldsymbol{x}_k, \boldsymbol{u}_k) + \alpha\left(h(\boldsymbol{x}_k)\right)] \ge \Gamma \Delta t^{r+1}.\end{aligned} \tag{30}$$

We then make $\Gamma\Delta t^{r+1} - w(\boldsymbol{x}_k, \boldsymbol{u}_k) \to \mathscr{W}(\boldsymbol{x}_k, \boldsymbol{u}_k, \sigma_{JU})$, which is a positively correlated function of σ_{JU}. As σ_{JU} increases, the constraint becomes stricter.

Furthermore, we calculate the percentile p of the current step's $\sigma_{JU,k}$ within the collected data and utilize ensemble critics to evaluate both $\boldsymbol{u}_k^{RL}$ and $\boldsymbol{u}_k^{CBF}$. $\boldsymbol{u}_k^{RL}$ is selected when the following condition is satisfying:

$$\mathbb{P}(CVaR(\boldsymbol{s}_k, \boldsymbol{u}_k^{RL}) \ge CVaR(\boldsymbol{s}_k, \boldsymbol{u}_k^{CBF})) > p, \tag{31}$$

where the probability is calculated as the proportion of ensemble critics that consider $\boldsymbol{u}_k^{RL}$ superior to $\boldsymbol{u}_k^{CBF}$, relative to the total number of critics.

IV. Implementations

A. Environment Settings

We implement a bidirectional four-lane intersection scenario based on Highway-Env [38]. Each SV is controlled by an improved Intelligent Driver Model (IDM) [39], which predicts its heading and position for the subsequent 2 s, yielding to potential collisions according to road priorities. When resetting the scenario, 10 SVs are initialized with random velocities between 6 m/s and 10 m/s. Concurrently, the EV is placed in a collision-free lane with a random velocity. Simulation frequency f_s is 15 Hz, with the policy execution frequency f_π set to 5 Hz during training and 10 Hz during testing.

B. MDPs Design

1) Observation Space and Action Space: The observation space includes the states of the EV $\mathcal{S}_{EV}$ and SVs $\mathcal{S}_{SV}$, the subsequent N_{wp} global reference waypoints $\mathcal{S}_{wp}$, and a one-hot task encoding $\mathcal{S}_{task}$ for left turn, go straight, or right turn: $\mathcal{S} = [\mathcal{S}_{EV}, \mathcal{S}_{SV}, \mathcal{S}_{wp}, \mathcal{S}_{task}]$, where S_{EV} = $[\mathbb{I}_{veh}$,x,y,v_x,v_y,φ,ω,d_{veh},d_{road},$d_{des}]$, x, y are the position coordinates of the vehicle's center of gravity, v_x, v_y are the longitudinal and lateral velocities, φ is the heading angle, ω is the yaw rate. $\mathbb{I}_{veh} \in \{0, 1\}$ is an indicator function to identify whether a vehicle is observed ($\mathbb{I}_{veh} = 1$). $d_{veh}, d_{road}, d_{des}$ are the distance to the closest vehicle, the road boundary, and the destination. $\mathcal{S}_{SV}$ contains N_{SV} SV's states: $\mathcal{S}_{SV}^j = [\mathbb{I}_{veh}$,$\Delta x$,$\Delta y$,$\Delta v_x$,$\Delta v_y$,$\varphi]$, $j \in [1, ..., N_{SV}]$. $\Delta x, \Delta y, \Delta v_x, \Delta v_y$ represent the position and velocity of the SVs relative to the EV. The waypoints in $\mathcal{S}_{wp}$ are also represented as the relative distances to the EV. The continuous action space is $\mathcal{A} = [a_{\text{lon}}, \delta_f]$.

2) Reward Design: The reward function consists of sparse $\mathbf{r}_{sparse}$ and dense $\mathbf{r}_{dense}$ rewards. Sparse rewards penalize collisions and encourage reaching target points:

$$\mathbf{r}_{sparse} = \mathbf{r}_{collision} + \mathbf{r}_{arrive\_goal}, \tag{32a}$$

$$\mathbf{r}_{collision} = -50 \cdot \mathbb{I}_{collision}, \tag{32b}$$

$$\mathbf{r}_{arrive\_goal} = 50 \cdot \mathbb{I}_{arrive\_goal}. \tag{32c}$$

The dense reward considers factors like reference line, action smoothness, distance to destination, and safety distance.

$$\mathbf{r}_{dense} = 3/(1 + \mathbf{r}_{ref}) + \mathbf{r}_{smooth} + \mathbf{r}_{des} + \mathbf{r}_{safe}, \tag{33a}$$

$$\mathbf{r}_{ref} = \max_{i=1,2} (\boldsymbol{x}_{k,i}^{\text{ref}} - \boldsymbol{x}_k)^\top Q (\boldsymbol{x}_{k,i}^{\text{ref}} - \boldsymbol{x}_k), \tag{33b}$$

$$\mathbf{r}_{act} = -(\boldsymbol{u}_k^\top R_{\boldsymbol{u}} \boldsymbol{u}_k + R_\Delta \Delta \boldsymbol{u}_k), \quad \mathbf{r}_{des} = -d_{des}^2, \tag{33c}$$

$$\mathbf{r}_{safe} = \begin{cases} 0.0 & d_{veh} > 5.0, \\ -(1.0 - d_{veh}) & 0.5 < d_{veh} \le 2.0, \\ -3 \times (1.0 - d_{veh}) & d_{veh} \le 0.5. \end{cases} \tag{33d}$$

where $\boldsymbol{x}_k^{\text{ref}} = [x^{\text{ref}}, y^{\text{ref}}, v_x^{\text{ref}}, 0, \phi^{\text{ref}}, 0]^\top$. For $\mathbf{r}_{ref}$, both reference lines are computed, and the maximum value is selected to guide the EV to stay close to the reference line. $\mathbf{r}_{act}$ is designed to promote energy efficiency and smooth the trajectory. The weight coefficients are $Q = diag(400.0, 400.0, 30.0, 30.0, 2.0, 0.5)$, $R_a = diag(0.05, 0.02)$, $R_\Delta = [0.2, 0.3]$.

C. Network Architecture and Hyper-Parameters

As shown in Fig. 3, the states of SVs, EV, waypoints, control signals, etc. are input into their respective encoding layers. For the waypoints, we apply cosine position encoding and employ learnable weights to capture the varying importance of each waypoint. Based on our previous work [40], we use a 2-hop attention structure to process the information of both SVs and the EV, thereby generating permutation-invariant features. The focus of this paper is not on the method for obtaining quantiles. Consequently, while fixed quantiles are utilized in the experiments, this does not

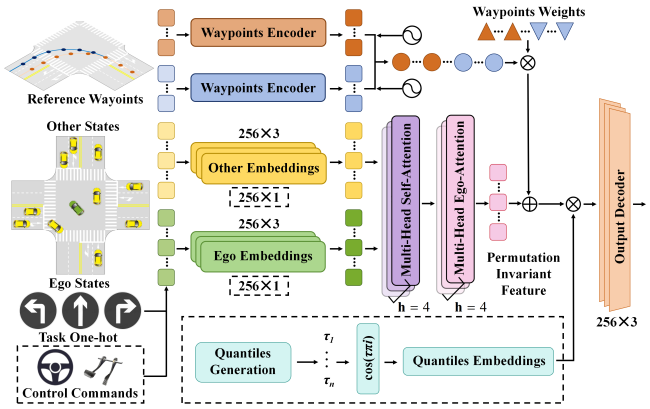

Fig. 3. Details of designed network. The components within the dashed box are used only in the critic network.

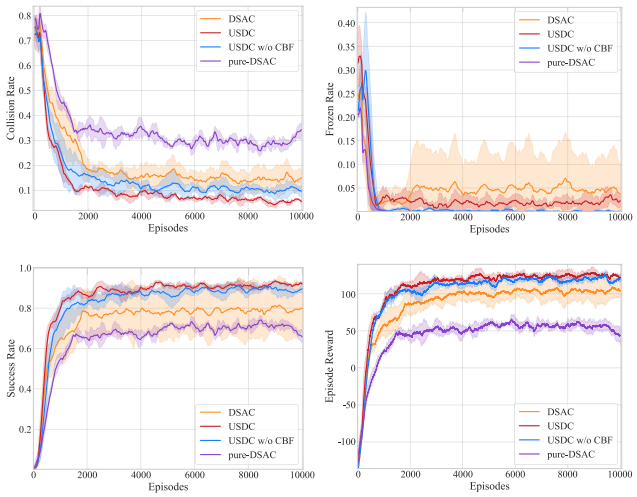

Fig. 4. Training curves of experiment. Solid lines correspond to the mean and the shaded regions correspond to 95% confidence interval over 3 runs.

preclude the use of more advanced techniques for quantile generation. The hyper-parameters used are listed in Table I.

TABLE I
HYPER-PARAMETERS

Hyper-parameter	Value	Hyper-parameter	Value
Network hidden size	256	Temperature factor	0.005
Activation function	GELU	Batch size	256
Actor learning rate α_π	3e-4→1e-5	Entropy learning rate	3e-4
Critic learning rate $\alpha_{r,c}$	3e-3→1e-4	Target entropy $\bar{\mathcal{H}}$	-dim($\mathcal{A}$)
Discount factor γ	0.99	Number of ensemble N	5
Prior factor ρ	10.0	Bernoulli mean p	0.9
σ_{AU}, σ_{EU}, σ_{JU}	10000	$CVaR$ risk parameter β	0.25
Learning buffer size	1e5	Number of quantile sample N_q	32
Vehicle class $\mathcal{K}$ function α_{veh}	$0.2x$	Number of closest points $\mathbf{N}$	5
Road class $\mathcal{K}$ function α_{road}	$0.5x$	TTCBF parameter Γ	300

D. Experiment Results

We compare USDC with the following baselines: distributional soft actor critic (DSAC [27]); pure-DSAC, which uses a similarly structured multi-layer perception (MLP) network for the network part; USDC without using HOCBF as a safety filter; and Dir-CBF, which directly applies HOCBF filtering to each action generated by a well-trained USDC agent without considering uncertainty. During training, each algorithm is trained three times with different random seeds. In each episode, tasks with randomly generated EVs are assigned, including left turn (LT), go straight (GS), and right turn (GT). During the testing phase, each algorithm is evaluated on 200 episodes for each of the three tasks.

The learning curves in Fig. 4 and the test results in Table II indicate that the USDC with a probabilistic safety filter improves performance while balancing traffic efficiency and safety. Despite an increase in FR compared to USDC without HOCBF, integrating the critic to estimate the $CVaR$ of $\boldsymbol{u}^{RL}$ and $\boldsymbol{u}^{CBF}$ provides safety guarantees under high uncertainty, resulting in a 11.0% reduction in CR compared to DSAC. In low uncertainty scenarios, the policy that maximizes $CVaR$ is selected by comprehensively assessing the long-term returns and instantaneous risks of $\boldsymbol{u}^{RL}$ and $\boldsymbol{u}^{CBF}$. This results in final performance metrics of 207.3 for AER and 7.77 m/s for AEV, matching or even surpassing other baselines. It is noteworthy that, although USDC increases FR by 3.5% compared to the USDC without HOCBF, its dynamic policy switching achieved through uncertainty quantification effectively avoids the drawbacks caused by over-conservatism in direct CBF. As shown in Table II, although the direct CBF method controls CR at 2.8%, its FR is as high as 27.8%. Additionally, the SR, AER, and AEV are reduced by 19.9%, 49.3%, and 12.6% respectively compared to USDC, indicating that relying solely on safety constraints may compromise the balance between efficiency and safety for AV.

TABLE II
COMPARE PERFORMANCE ON THREE DRIVING TASKS.

Tasks	Algorithms	SR(%)	FR(%)	CR(%)	AER	AEV(m/s)
LT	Pure-DSAC	56.5	**0.5**	43.0	-22.7±48.2	**7.99±0.64**
	DSAC	82.5	1.0	19.5	100.9±47.5	7.00±1.33
	Dir-CBF	72.0	28.0	**4.5**	6.1±64.0	6.34±1.51
	USDC (ours)	**91.0**	5.0	6.0	**129.9±47.1**	7.71±0.44
	USDC w/o CBF	86.5	**0.5**	15.5	117.4±46.8	7.91±0.45
GS	Pure-DSAC	62.5	3.5	34.0	30.8±52.4	**7.85±0.74**
	DSAC	73.0	7.5	23.0	147.2±53.4	7.73±0.53
	Dir-CBF	56.0	48.0	**3.0**	58.8±69.5	6.81±1.04
	USDC (ours)	**88.5**	9.0	5.0	**182.7±47.7**	7.59±0.37
	USDC w/o CBF	83.5	**3.0**	14.5	172.4±49.5	7.57±0.43
RT	Pure-DSAC	88.5	0.5	11.5	190.8±30.1	7.98±0.57
	DSAC	98.0	0.5	1.5	267.8±19.5	**8.06±0.42**
	Dir-CBF	92.0	7.5	1.0	232.0±37.1	7.21±0.87
	USDC (ours)	**100.0**	**0.0**	**0.0**	**309.2±12.1**	8.01±0.22
	USDC w/o CBF	98.5	**0.0**	1.5	278.9±17.1	7.99±0.29
MEAN	Pure-DSAC	69.2	1.5	30.2	66.3±43.6	**7.94±0.65**
	DSAC	84.5	3.0	14.7	172.0±40.1	7.60±0.76
	Dir-CBF	73.3	27.8	**2.8**	99.0±56.8	6.79±1.14
	USDC (ours)	**93.2**	4.7	3.7	**207.3±35.6**	7.77±0.34
	USDC w/o CBF	89.5	**1.2**	10.5	189.6±37.8	7.82±0.39

[1] SR, FR, CR, AER, and AEV stand for success rate, frozen rate, collision rate, average episode reward, and average episode velocity, respectively. The definitions are similar to those in [40], except that FR is defined as a runtime exceeding 20 seconds (200 steps).
[2] Bold: best performance. Policy update frequency $f_\pi = 10$ Hz.

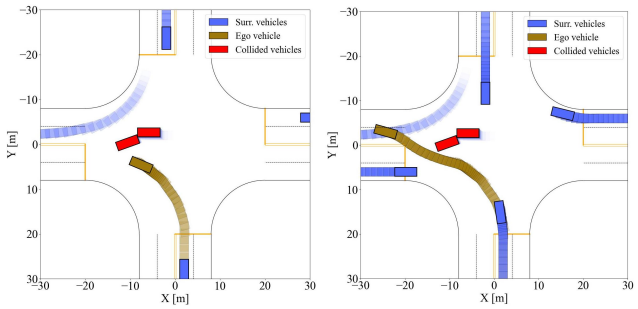

Fig. 5. Vehicle trajectories visualization. Blue rectangles are SVs, brown rectangles are EVs. Red are collided vehicles. (a) 124 steps. (b) 184 steps.

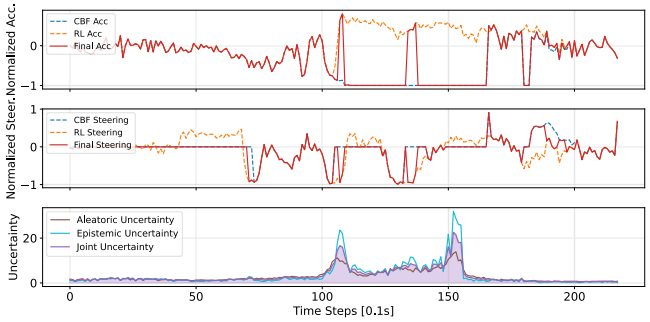

Fig. 6. Vehicle control commands and uncertainty during the case.

E. Case Study

In the real world, intersections are accident prone scenarios and the AV still needs to make safe and robust decisions when it encounters the presence of accident vehicles at the intersection. Fig. 5 demonstrates a left turn unsignalized intersection scenario where two SVs collide. Such scenarios are rare in both the real world and during training, making them an OOD challenge for RL decision making. As shown in Fig. 6, the AV approaches the accident vehicle at around 100 steps, leading to a significant increase in decision difficulty due to compressed passage space. Meanwhile, both EU and AU increase rapidly, indicating increased scenario randomness and decreased reliability of RL policy. Therefore, the AV needs to apply the deceleration braking advice given by the safety filter to avoid the potential collision risk. After gradually passing the obstacle at low speed, the AV's uncertainty decreases. At around 180 steps, due to the close distance between the AV and the left road boundary, the safety filter generates a corrective command to turn right. The AV then evaluates the original policy output $\boldsymbol{u}^{RL}$ and the corrective command $\boldsymbol{u}^{CBF}$, gradually adjusting the control commands to move away from the road boundary. The results show that the USDC can effectively identify the above OOD scenario and enable the AV to safely and reliably pass through the scenario.

V. Conclusions

In this paper, we propose an uncertainty-aware safety-critical decision and control framework. USDC generates a risk-averse policy by constructing a risk-aware distributional ensemble RL, while estimating uncertainty to quantify the reliability of the policy. FPN and Bootstrapping are employed to enhance the diversity of critics when encountering safety-critical scenarios. Subsequently, a HOCBF is employed as a safety filter to minimize intervention policy, while dynamically enhancing constraints based on uncertainty. Instead of relying on a specific threshold, the ensemble critics evaluate both HOCBF and RL policies simultaneously based on the JU distribution. This ensures that the RL policy performs no worse than the HOCBF under low uncertainty and favors safer policies under high uncertainty. Experimental results for left turn, right turn, and go straight driving tasks demonstrate that USDC improves safety compared to baseline methods while maintaining traffic efficiency. The forward invariance of the HOCBF may be violated because of input constraints. Additionally, the HOCBF in this paper only guarantees safety for a single step and does not account for the effects of changes in obstacle states. Future work can be extended to a more robust safety filter to improve overall safety.

References

[1] H. Wang, W. Shao, C. Sun, K. Yang, D. Cao, and J. Li, "A survey on an emerging safety challenge for autonomous vehicles: Safety of the intended functionality," *Engineering*, vol. 33, pp. 17–34, 2024.

[2] M. Abdel-Aty and S. Ding, "A matched case-control analysis of autonomous vs human-driven vehicle accidents," *Nat. Commun.*, vol. 15, no. 1, p. 4931, 2024.

[3] S. Noh, "Decision-making framework for autonomous driving at road intersections: Safeguarding against collision, overly conservative behavior, and violation vehicles," *IEEE Trans. Ind. Electron.*, vol. 66, no. 4, pp. 3275–3286, 2019.

[4] M. T. Ashraf, K. Dey, S. Mishra, and M. T. Rahman, "Extracting rules from autonomousvehicle-involved crashes by applying decision tree and association rule methods," *Transp. Res. Rec.*, vol. 2675, no. 11, pp. 522 – 533, 2021.

[5] Z. Li, L. Xiong, B. Leng, P. Xu, and Z. Fu, "Safe reinforcement learning of lane change decision making with risk-fused constraint," in *Proc. IEEE Intell. Transp. Syst. Conf.*, 2023, pp. 1313–1319.

[6] G. Li, W. Zhou, S. Lin, S. Li, and X. Qu, "On-ramp merging for highway autonomous driving: An application of a new safety indicator in deep reinforcement learning," *Automot. Innov.*, vol. 6, no. 3, pp. 453 – 465, 2023.

[7] H. Hu, D. Chu, J. Yin, and L. Lu, "Double deep q-networks based game-theoretic equilibrium control of automated vehicles at autonomous intersection," *Automot. Innov.*, vol. 7, no. 4, pp. 571 – 587, 2024.

[8] K. Yang, S. Li, Y. Chen, D. Cao, and X. Tang, "Towards safe decision-making for autonomous vehicles at unsignalized intersections," *IEEE Trans. Veh. Technol.*, vol. 74, no. 3, pp. 3830–3842, 2025.

[9] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, "A review of safe reinforcement learning: Methods, theories, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 11 216–11 235, 2024.

[10] A. Stooke, J. Achiam, and P. Abbeel, "Responsive safety in reinforcement learning by pid lagrangian methods," in *Int. Conf. Mach. Learn.* PMLR, 2020, pp. 9133–9143.

[11] H. Honari, A. M. S. Enayati, M. G. Tamizi, and H. Najjaran, "Meta sac-lag: Towards deployable safe reinforcement learning via metagradient-based hyperparameter tuning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.* IEEE, 2024, pp. 619–626.

[12] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Int. Conf. Mach. Learn.* PMLR, 2017, pp. 22–31.

[13] Y. Zhang, Q. Vuong, and K. Ross, "First order constrained optimization in policy space," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 15 338–15 349, 2020.

[14] Z. Shixin, P. Feng, J. Anni, Z. Hao, and G. Qiuqi, "The unmanned vehicle on-ramp merging model based on am-mappo algorithm," *Sci. Rep.*, vol. 14, no. 1, p. 19416, 2024.

[15] H. Ma, C. Liu, S. E. Li, S. Zheng, W. Sun, and J. Chen, "Learn zero-constraint-violation safe policy in model-free constrained reinforcement learning," *IEEE Trans. Neural Netw. Learn, Syst.*, vol. 36, no. 2, pp. 2327–2341, 2025.

[16] H. Zheng, H. Ma, S. Zheng, S. E. Li, and J. Wang, "Synthesize efficient safety certificates for learning-based safe control using magnitude regularization," in *IEEE Int. Conf. Robot. Autom.*, 2024, pp. 545–551.

[17] X. Wang, "Ensuring safety of learning-based motion planners using control barrier functions," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4773–4780, 2022.

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

[19] W. Zhou, Z. Cao, N. Deng, K. Jiang, and D. Yang, "Identify, estimate and bound the uncertainty of reinforcement learning for autonomous driving," *IEEE Trans. Intell. Transport. Syst.*, vol. 24, no. 8, pp. 7932–7942, 2023.

[20] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Mach. Learn.*, vol. 110, no. 3, pp. 457–506, 2021.

[21] S. Wang and S. Wen, "Safe control against uncertainty: A comprehensive review of control barrier function strategies," *IEEE Syst. Man Cybern. Mag.*, vol. 11, no. 1, pp. 34–47, 2025.

[22] C.-J. Hoel, K. Wolff, and L. Laine, "Ensemble quantile networks: Uncertainty-aware reinforcement learning with applications in autonomous driving," *IEEE Trans. Intell. Transport. Syst.*, vol. 24, no. 6, pp. 6030–6041, 2023.

[23] Z. Zhang, Q. Liu, Y. Li, K. Lin, and L. Li, "Safe reinforcement learning in autonomous driving with epistemic uncertainty estimation," *IEEE Trans. Intell. Transport. Syst.*, vol. 25, no. 10, pp. 13 653–13 666, 2024.

[24] K. Yang, X. Tang, S. Qiu, S. Jin, Z. Wei, and H. Wang, "Towards robust decision-making for autonomous driving on highway," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 11 251–11 263, 2023.

[25] X. Tang, G. Zhong, S. Li, K. Yang, K. Shu, D. Cao, and X. Lin, "Uncertainty-aware decision-making for autonomous driving at uncontrolled intersections," *IEEE Trans. Intell. Transport. Syst.*, vol. 24, no. 9, pp. 9725–9735, 2023.

[26] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, p. 105151, 2022.

[27] X. Ma, L. Xia, Z. Zhou, J. Yang, and Q. Zhao, "Dsac: Distributional soft actor critic for risk-sensitive reinforcement learning," *arXiv preprint arXiv:2004.14547*, 2020.

[28] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

[29] W. Xiao and C. Belta, "High-order control barrier functions," *IEEE Trans. Autom. Control.*, vol. 67, no. 7, pp. 3655–3662, 2022.

[30] Y. Xiong, D.-H. Zhai, M. Tavakoli, and Y. Xia, "Discrete-time control barrier function: High-order case and adaptive case," *IEEE Trans. Cybern.*, vol. 53, no. 5, pp. 3231–3239, 2023.

[31] R. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *J. Bank. Financ.*, vol. 26, no. 7, pp. 1443–1471, 2002.

[32] A. Rame and M. Cord, "Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation," in *Proc. Int. Conf. Learn. Represent.*, 2021.

[33] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[34] I. Osband, J. Aslanides, and A. Cassirer, "Randomized prior functions for deep reinforcement learning," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[35] R. Rajamani, *Vehicle dynamics and control*. Springer Science & Business Media, 2011.

[36] J. Xu and B. Alrifaee, "High-order control barrier functions: Insights and a truncated taylor-based formulation," *arXiv preprint arXiv:2503.15014*, 2025.

[37] J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl, "CasADi – A software framework for nonlinear optimization and optimal control," *Mathematical Programming Computation*, 2018.

[38] E. Leurent, "An environment for autonomous driving decision-making," https://github.com/eleurent/highway-env, 2018.

[39] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Phys. Rev. E*, pp. 1805–1824, Jul 2002.

[40] B. Leng, R. Yu, W. Han, L. Xiong, Z. Li, and H. Huang, "Risk-aware reinforcement learning for autonomous driving: Improving safety when driving through intersection," *arXiv preprint arXiv:2503.19690*, 2025.