Version Age of Information Oriented Risk-Sensitive Scheduling with Distributional Reinforcement Learning

Chen Chen*, Haoyuan Pan*, and Tse-Tin Chan†
*College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
†Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong, China
E-mails: 2400101055@mails.szu.edu.cn, hypan@szu.edu.cn, tsetinchan@eduhk.hk

***Abstract*—Timely delivery of fresh information is critical in real-time wireless systems, where outdated updates can compromise performance or safety. The recently proposed Version Age of Information (VAoI) extends the classical Age of Information (AoI) by measuring the version gap between transmitted and received updates, capturing freshness in version-asynchronous environments. Existing works primarily optimize average AoI or VAoI, overlooking rare but catastrophic staleness events. To address this gap, we study a risk-sensitive VAoI scheduling problem in multiuser wireless networks under resource constraints, formulated as a Constrained Markov Decision Process (CMDP). Unlike conventional reinforcement learning, which estimates only expected returns, our approach leverages distributional DRL (C51, QR-DQN, IQN) to model the full return distribution, enabling optimization of the Conditional Value at Risk (CVaR), a principled risk measure that quantifies extreme VAoI realizations beyond the mean. We design a scheduling framework that minimizes CVaR of VAoI while respecting resource budgets. Simulation results demonstrate that the proposed method not only maintains competitive average VAoI but also significantly reduces the CVaR of VAoI in IQN-based implementations, providing a flexible trade-off between average performance and tail-risk mitigation. These results highlight the essential role of distributional learning for robust, low-risk status update scheduling in wireless systems.**

***Index Terms*—Version age of information, risk-sensitive, distributional reinforcement learning, conditional value at risk.**

I. Introduction

Timely access to fresh information is critical in real-time and safety-sensitive applications such as industrial automation and intelligent transportation. Delays in delivering fresh updates can degrade control performance and, in extreme cases, compromise system safety. To quantify information freshness, the Age of Information (AoI) has become a widely adopted metric, measuring the time elapsed since the most recently generated update was successfully received [1].

While AoI offers a time-based view of staleness, it does not always capture the true relevance of information. In many systems, updates arrive in successive versions, and not every received update meaningfully advances the system state. AoI resets upon reception of any update, even if the new packet carries negligible semantic value. To address this limitation, the Version Age of Information (VAoI) has been introduced [2], [3]. VAoI explicitly measures the version gap between transmitter and receiver, reflecting how far behind the receiver is in terms of version progression. This distinction is particularly critical in software-driven and data-centric applications, such as wireless control systems and industrial automation, where synchronization with the latest version directly impacts safety and performance. In such environments, VAoI provides a more appropriate freshness metric than AoI by aligning update importance with version consistency rather than elapsed time alone.

Despite growing attention to AoI and VAoI optimization, most existing studies adopt risk-neutral formulations, focusing solely on minimizing long-term average performance [4], [5]. However, strategies optimized for average VAoI may still suffer from rare but catastrophic spikes of extreme staleness. In safety-critical systems, a single episode of severe information lag can destabilize control or even trigger system failure, regardless of otherwise good average performance. This underscores the need to explicitly control *tail risks* rather than averages alone. Traditional model-based optimization methods [6], [7] are further limited by their reliance on stationary dynamics, which may not hold in practical wireless environments. Meanwhile, conventional deep reinforcement learning (DRL), though model-free, typically optimizes expected returns only, overlooking return variability and failing to capture tail behavior in VAoI.

To explicitly account for such risks, we adopt Conditional Value at Risk (CVaR) [8] as a principled performance criterion. CVaR measures the expected value of losses (e.g., VAoI in our context) beyond a given quantile, capturing the worst-case tail performance while maintaining desirable mathematical properties such as convexity and subadditivity. Optimizing CVaR, however, fundamentally requires access to the entire return distribution. This motivates using distributional deep reinforcement learning (distributional DRL), which estimates the full distribution of returns rather than a single expectation value. Recent studies have begun to explore distributional DRL in wireless communication and networking, where stochastic dynamics and uncertainty are intrinsic. Applications include wireless resource allocation and radio management, with evidence of improved robustness over expectation-based DRL [9], [10]. Unlike conventional DRL, which optimizes only average performance, distributional DRL methods such as Categorical DQN (C51) [11], Quantile Regression DQN (QR-DQN) [12], and Implicit Quantile Networks (IQN) [13] explicitly model

This work was supported in part by the National Natural Science Foundation of China under Grant 62371302, in part by Shenzhen Science and Technology Program under Grant JCYJ20250604181549064, and in part by the Research Matching Grant Scheme from the Research Grants Council of Hong Kong. (*Corresponding author: Haoyuan Pan.*)

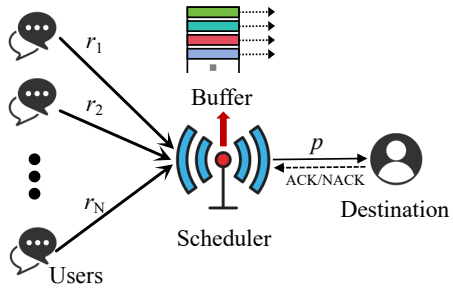

Fig. 1. System model: a scheduler receives data packets from N independent users and forwards them to a destination over a wireless channel.

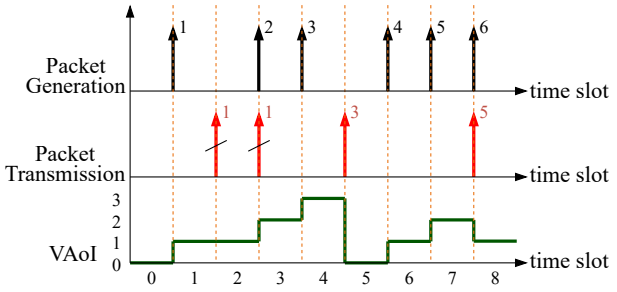

Fig. 2. Example of VAoI evolution for a single user. Black arrows denote packet generation at the user, red arrows indicate successful transmissions to the destination, and a black slash marks transmission failure.

variability in returns, thereby enabling direct optimization of tail-sensitive objectives such as CVaR.

This paper studies a risk-sensitive VAoI scheduling problem in multiuser wireless networks under transmission resource constraints, formulating it as a Constrained Markov Decision Process (CMDP) [14]. To solve this problem in a model-free yet risk-aware manner, we develop a distributional DRL framework that leverages return distributions for CVaR minimization, enabling fine-grained risk control. Specifically, we investigate multiple distributional DRL algorithms, including C51, QR-DQN, and IQN, to approximate return distributions and derive CVaR-based scheduling policies at different risk levels. In addition, we design a tunable utility function that allows system designers to flexibly trade off between minimizing average VAoI and mitigating extreme staleness.

Simulation results demonstrate that the proposed distributional DRL methods substantially outperform baseline non-distributional strategies in mitigating tail risks, as measured by the CVaR of VAoI. In particular, the IQN-based approach achieves up to a 28.5% reduction in CVaR compared with non-distributional methods. Moreover, our framework enables—for the first time in VAoI scheduling, a flexible and interpretable trade-off between average performance and tail-risk mitigation, allowing system designers to balance efficiency and reliability according to application requirements.

II. System Model and Risk-Neutral Scheduling

A. System Model

We consider a discrete-time multiuser update system, illustrated in Fig. 1. A central scheduler collects status updates from N independent users and forwards them to a destination over a shared wireless channel. Time is slotted, and at the beginning of each slot t, user $n \in \{1, 2, \ldots, N\}$ generates a new packet with probability $r_n \in (0, 1]$. Newly generated packets are delivered instantaneously and reliably to the scheduler, which maintains a buffer of size N, storing only the most recent packet of each user.

Due to limited spectrum resources, the scheduler can transmit at most one packet per time slot. Specifically, at each time slot t, the scheduling action is $a_t \in \mathcal{A} = \{0, 1, \ldots, N\}$, where $a_t = 0$ denotes no transmission and $a_t = n$ denotes transmitting the latest packet of user n. Each scheduled transmission succeeds with probability $p \in (0, 1]$. At the end of the slot, the destination returns an acknowledgment (ACK) or negative acknowledgment (NACK) over a reliable feedback channel to indicate transmission success or failure.

To quantify the freshness of information, we adopt the VAoI metric. Unlike classical AoI, which measures the time elapsed since the last update was received, VAoI tracks semantic staleness by monitoring the version gap between the scheduler and destination. Let $G(t, n)$ denote the version index of user n's packet stored at the scheduler at the beginning of slot t. If a new packet arrives, $G(t, n) = G(t-1, n) + 1$; otherwise, $G(t, n) = G(t-1, n)$. Let $B(t, n)$ denote the version index of the latest packet from user n successfully received by the destination at the end of slot t. If a packet from user n is transmitted and successfully delivered, then $B(t, n) = G(t, n)$; otherwise, $B(t, n) = B(t-1, n)$. Therefore, the instantaneous VAoI of user n at the beginning of slot t is defined as

$$\mathcal{D}(t, n) = G(t, n) - B(t-1, n). \tag{1}$$

Fig. 2 illustrates an example of the instantaneous VAoI evolution for a single user. Black arrows denote packet generation at the user and its delivery to the central scheduler. Red arrows indicate successful transmissions to the destination, and a black slash marks transmission failure. The version number of a packet is marked next to each arrow. As shown in the figure, a new packet is generated at the beginning of slot 1, raising the VAoI to 1. Next, since the scheduled transmission in slot 1 fails and no new packet is generated at slot 2, the VAoI remains unchanged at the beginning of slot 2. The VAoI continues to accumulate, when the version gap reaches 3 at the beginning of slot 5. At this point, a successful transmission occurs, and the VAoI resets to 0 (and no new packet is generated). Later, at the end of slot 7, a packet with version 5 is successfully delivered; however, because a newer packet with version 6 is generated at the beginning of slot 8, the VAoI is updated to $1 = (6-5)$. This example highlights how VAoI depends jointly on packet generation, transmission outcomes, and version progression, rather than on elapsed time alone.

B. Risk-Neutral Scheduling

Most prior works minimize the long-term average system-wide VAoI, expressed as

$$\bar{\mathcal{D}} = \lim_{T \to \infty} \frac{1}{TN} \sum_{t=1}^{T} \sum_{n=1}^{N} \mathcal{D}(t, n), \tag{2}$$

where T is the horizon length. This is a risk-neutral objective that emphasizes expected performance. In many applications, the scheduler should satisfy a long-term transmission cost constraint. The average transmission cost is defined as

$$\eta = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}[a_t \neq 0], \tag{3}$$

where $\mathbb{1}[\cdot]$ is the indicator function that returns 1 when its argument is true. The scheduler must ensure $\eta \leq \eta_{\max}$, where $\eta_{\max} \in (0, 1]$ is the maximum average transmission cost budget. The risk-neutral scheduling problem is formulated as

$$\begin{aligned} &\min\nolimits_{\{a_1,\ldots,a_T\}} \ \bar{\mathcal{D}} \\ &\text{s.t. } \eta \leq \eta_{\max} \end{aligned} \tag{4}$$

This problem can be modeled as a CMDP. The action space is defined by $a_t \in \mathcal{A} = \{0, 1, \ldots, N\}$ as above. The state at time t is the VAoI vector $s_t = (\mathcal{D}(t, 1), \mathcal{D}(t, 2), \ldots, \mathcal{D}(t, N))$, with each $\mathcal{D}(t, n) \in \{0, 1, \ldots, \mathcal{D}_{\max}\}$. The truncation $\mathcal{D}_{\max}$ ensures a finite state space, motivated by (i) in practice, packets lose value beyond a freshness threshold, and (ii) unbounded VAoI complicates analysis without adding further insight.

State transition dynamics $P(s_{t+1}|s_t, a_t)$ defining distribution over the next state s_{t+1}, conditioned on the current state s_t and action a_t, depend on packet generation rates r_n and channel success probability p. The instantaneous cost function c_t is determined by the action a_t at time slot t: if $a_t = 0$, then $c_t = 0$; otherwise, $c_t = 1$. The reward function is defined to provide a scalar reward for executing action a_t in state s_t. The immediate reward at time t is defined as

$$r_t = -\sum_{n=1}^{N} \mathcal{D}(t, n) - \lambda c_t, \tag{5}$$

where the first term penalizes large VAoI values, and the second term enforces the transmission constraint through the Lagrange multiplier $\lambda \geq 0$. The multiplier is updated as $\lambda \leftarrow \max(0, \lambda + \delta \cdot (\eta - \eta_{\max}))$ with step size δ, where η is the empirical average cost. This update follows standard dual optimization principles in CMDPs and is widely used in CMDP-based reinforcement learning [14]. A discount factor $\gamma \in [0, 1]$ balances short- and long-term rewards, trading off immediate VAoI reduction against future performance.

Within this CMDP framework, conventional model-free DRL methods such as DQN can effectively learn scheduling policies that minimize the expected long-term VAoI while satisfying resource constraints. However, such approaches remain risk-neutral: they optimize average freshness but fail to account for the distributional tail of VAoI, where rare but severe staleness events occur. Addressing these extreme events requires reformulating the problem with risk-sensitive objectives and leveraging distributional DRL methods, as discussed in the next section.

III. Risk-Sensitive Scheduling with Distributional DRL

This section presents a risk-sensitive scheduling scheme for minimizing VAoI in multiuser wireless systems. We first reformulate the original scheduling problem using CVaR-based risk criteria. We then design a distributional DRL framework that leverages the full return distribution for tail-aware scheduling.

A. Risk-Sensitive Scheduling Problem Formulation

We focus on the tail performance of system-wide VAoI by adopting the CVaR of its empirical distribution. Specifically, we construct a mixed distribution by pooling all VAoI samples across users and time slots, i.e.,

$$\mathcal{Z} = \{\mathcal{D}(t, n) : t = 1, \ldots, T,\ n = 1, \ldots, N\}. \tag{6}$$

Let $F_{\mathcal{Z}}$ denote the empirical cumulative distribution function (CDF) of $\mathcal{Z}$. For a risk level $\alpha \in (0, 1)$, the mixed-distribution CVaR of VAoI in the multiuser system is defined as

$$\text{CVaR}_\alpha(\mathcal{Z}) = \min_{z\in\mathbb{R}} \left\{ z + \frac{1}{\alpha TN} \sum_{t=1}^{T} \sum_{n=1}^{N} \left[\mathcal{D}(t, n) - z\right]_+ \right\}, \tag{7}$$

where $[x]_+ = \max\{x, 0\}$. This metric captures the average of the worst α fraction of VAoI samples across all users and time slots, thereby characterizing the system-wide tail behavior. The risk-sensitive scheduling problem is thus formulated as

$$\begin{aligned} &\min\nolimits_{\{a_1,\ldots,a_T\}} \ \text{CVaR}_\alpha(\mathcal{Z}) \\ &\text{s.t. } \eta \leq \eta_{\max} \end{aligned} \tag{8}$$

In conventional DRL, the per-slot reward r_t is designed so that maximizing the expected cumulative return minimizes average VAoI while satisfying the transmission cost. Standard methods, such as DQN, approximate the state-action value function $Q^\pi(s, a)$, representing the expected cumulative reward under policy π. However, they do not capture rare but extreme spikes in VAoI that may compromise safety or reliability. To address this, we adopt a distributional DRL approach, which models the full return distribution rather than its expectation, enabling risk-sensitive optimization using quantile-based measures such as CVaR.

B. CVaR-Aware Distributional DRL Design

Unlike conventional DQN, which models the return by its expectation, distributional DRL represents the return as a random variable $Z^\pi(s, a)$, thereby capturing both average performance and tail risks. This naturally provides a foundation for CVaR-based optimization.

Early distributional DRL methods include C51 and QR-DQN. C51 represents $Z^\pi(s, a)$ by a categorical distribution over a fixed set of discrete atoms, estimating only their probabilities. While simple, its accuracy depends heavily on the number of atoms and a predefined value range. QR-DQN eliminates this limitation by directly approximating the quantile function $F^{-1}_{Z^\pi}(\tau)$ at $\mathcal{N}$ fixed quantile levels $\{\hat{\tau}_i\}$, with a neural network outputting quantile estimates $\{F^{-1}_{Z^\pi}(\hat{\tau}_i)\}$. The expected return is then obtained as the average of these estimates, i.e., $\mathbb{E}[Z] = \frac{1}{\mathcal{N}} \sum_i F^{-1}_Z(\hat{\tau}_i)$. However, its fidelity is limited by the choice and density of quantile locations.

To overcome these limitations, Implicit Quantile Networks (IQN) generalize QR-DQN by learning the entire quantile

function $F^{-1}_{Z^\pi}(\tau|s,a)$ for any $\tau \in [0,1]$. During training, quantile samples τ are drawn from a uniform distribution, allowing the network to flexibly approximate the full return distribution and capture extreme tail events with higher accuracy. Given its expressive power, we adopt IQN as the backbone for explaining the risk-sensitive VAoI scheduling below. While our framework focuses on IQN, it is compatible with other distributional DRL algorithms, and comparative results are provided in Section IV.

1) IQN Architecture and Training: In IQN, the system state $s_t = (\mathcal{D}(t,1),\dots,\mathcal{D}(t,N))$ is first mapped into a feature vector $\psi(s_t)$ via a Multi-Layer Perceptron (MLP). Each sampled quantile $\tau \sim U([0,1])$ is encoded by an embedding network $\phi(\tau)$, constructed using cosine basis functions, to match the dimension of $\psi(s_t)$. This embedding allows the network to produce a return estimate specifically for that quantile, effectively conditioning the output on τ. In other words, the network can predict the value corresponding to any point in the return distribution rather than only the mean. The fused representation $\psi(s_t) \odot \phi(\tau)$, where $\odot$ denotes element-wise multiplication, is then processed by another MLP to output quantile estimates $Z^\pi_\tau(s,a)$ for all actions.

Training minimizes the quantile regression Huber loss. For a transition (s_t, a_t, r_t, s_{t+1}), we sample $\mathcal{N}$ quantiles $\{\tau_i\}_{i=1}^{\mathcal{N}}$ for (s_t,a_t) and $\mathcal{N}'$ quantiles $\{\tau_j\}_{j=1}^{\mathcal{N}'}$ for (s_{t+1},a') (often $N = N'$). Given samples τ_i and τ_j, the distributional version of temporal difference (TD) error is defined as

$$u_t^{i,j} = r_t + \gamma Z_{\tau_j}(s_{t+1},a') - Z_{\tau_i}(s_t,a_t), \tag{9}$$

where a' is the next action, and the risk-sensitive action selection policy π^{risk} is explained later. The IQN loss is then

$$\mathcal{L}(s,a,r,s') = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \sum_{j=1}^{\mathcal{N}} \rho^\kappa_{\tau_i}(u_t^{i,j}), \tag{10}$$

where $\rho^\kappa_\tau(\cdot)$ denotes the quantile regression Huber loss

$$\rho^\kappa_{\tau_i}(u) = \begin{cases} |\tau_i - \mathbb{I}\{u<0\}| \cdot \frac{1}{2}u^2, & \text{if } |u| \le \kappa, \\ |\tau_i - \mathbb{I}\{u<0\}| \cdot \kappa(|u| - \frac{1}{2}\kappa), & \text{otherwise.} \end{cases} \tag{11}$$

By iteratively minimizing this loss, IQN learns to approximate the full return distribution. The update rule for the network parameters θ follows $\theta \leftarrow \theta - \mu \nabla_\theta \mathcal{L}$, where μ is the learning rate and $\nabla_\theta \mathcal{L}$ is the gradient of the loss with respect to θ. To stabilize training, a target network with parameters θ^- is periodically synchronized with the primary network ($\theta^- \leftarrow \theta$), providing fixed targets for TD updates and reducing instability caused by rapidly fluctuating predictions.

2) Risk-Sensitive Decision-Making Based on CVaR: Given the learned distribution $Z^\pi(s,a)$, risk-sensitive measures of the return, such as Value at Risk (VaR) and CVaR, can be computed for risk level $\varphi \in [0,1]$:

$$\text{VaR}_\varphi[Z^\pi(s,a)] = \inf\{z \in \mathbb{R} : F_{Z^\pi}(z) \ge \varphi\}, \tag{12}$$

$$\text{CVaR}_\varphi[Z^\pi(s,a)] = \mathbb{E}[z \mid z \le \text{VaR}_\varphi[Z^\pi(s,a)]], \tag{13}$$

Algorithm 1 Risk-Sensitive Distributional DRL Scheduling

Initialize replay buffer $\mathcal{M}$; network parameters θ and target network θ^-; Lagrange multiplier $\lambda \ge 0$; empirical average cost $\eta = 0$; risk-tradeoff parameter β; CVaR risk level φ.
for each time step $t = 0, 1, \dots$ **do**
  Observe current state s_t
  (Estimate distributional quantities for decision)
  Sample K_{sel} quantiles $\{\tau_k\}_{k=1}^{K_{\text{sel}}} \sim U([0,1])$;
  For all $a \in \mathcal{A}$: (i) Compute quantile estimates $\{Z_{\tau_k}(s_t,a)\}_{k=1}^{K_{\text{sel}}}$ via IQN; (ii) Estimate expected return $Q^\pi(s_t,a)$ using (15); (iii) Estimate $\text{CVaR}_\varphi[Z^\pi(s_t,a)]$ by averaging quantiles $\tau \le \varphi$; (iv) Compute utility based on (14) for all $a \in \mathcal{A}$;
  Select action $a_t = \arg\max_{a\in\mathcal{A}} \mathcal{U}^\pi(s_t,a)$ and use ϵ-greedy for exploration;
  Execute a_t, observe reward r_t, next state s_{t+1}, and cost $c_t = \mathbb{1}[a_t \ne 0]$;
  Store transition (s_t, a_t, r_t, s_{t+1}) in $\mathcal{M}$;
  Update empirical average cost estimate: $\eta \leftarrow ((t-1)\eta + c_t)/t$.
  (Learning update)
  Sample minibatch of transitions from $\mathcal{M}$
  Sample $\mathcal{N}$ quantiles for current states and $\mathcal{N}'$ quantiles for next states;
  Compute target quantile distribution using target network θ^-;
  Compute quantile Huber loss $\mathcal{L}$ based on (10);
  Update network parameters: $\theta \leftarrow \theta - \mu \nabla_\theta \mathcal{L}$;
  Periodically update target network: $\theta^- \leftarrow \theta$;
  Update Lagrange multiplier for transmission cost: $\lambda \leftarrow \max\{0, \lambda + \delta(\eta - \eta_{\max})\}$.
end for

where $F_{Z^\pi}(z)$ is the CDF of $Z^\pi(s,a)$. $\text{CVaR}_\varphi[Z^\pi(s,a)]$ can be estimated efficiently by averaging quantile samples over the worst $\tau \in [0,\varphi]$ in IQN.

While our ultimate goal is to minimize the CVaR of the long-term system-wide mixed VAoI distribution $\text{CVaR}_\alpha(\mathcal{Z})$ in (7), directly optimizing this objective is computationally challenging. Evaluating the CVaR of the mixed-distribution VAoI involves tail risk analysis over multiple interdependent random variables, which becomes prohibitively expensive in high-dimensional or multiuser scenarios. To address this, we instead optimize the CVaR of the return distribution, $\text{CVaR}_\varphi[Z^\pi(s,a)]$. Recall that the reward in (5) is defined as the negative sum of instantaneous VAoI with a possible transmission cost penalty; hence, the left tail of the return distribution corresponds to the right tail of the cumulative VAoI distribution. Maximizing $\text{CVaR}_\varphi[Z^\pi(s,a)]$ therefore effectively minimizes worst-case VAoI scenarios. Although this surrogate does not guarantee exact minimization of $\text{CVaR}_\alpha(\mathcal{Z})$, it provides a tractable and meaningful approximation when computational tractability is a primary concern. In practice, we set the CVaR risk level φ equal to α.

In general, to balance average performance with risk sensi-

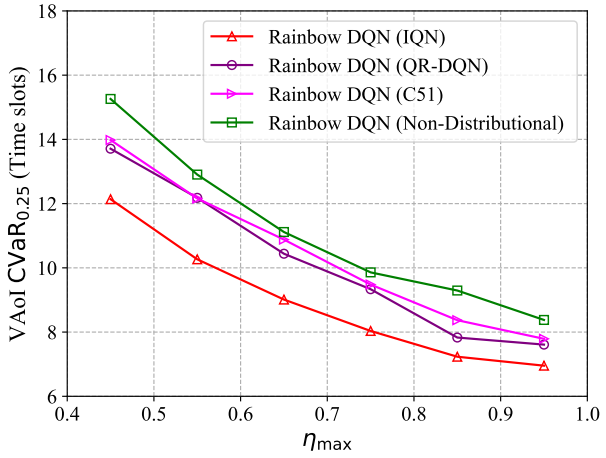

Fig. 3. The $\mathrm{CVaR}_{0.25}$ of system-wide VAoI versus $\eta_{\max}$ under different DRL algorithms built on the Rainbow framework. $p = 0.8$ and $r_n = 0.7$.

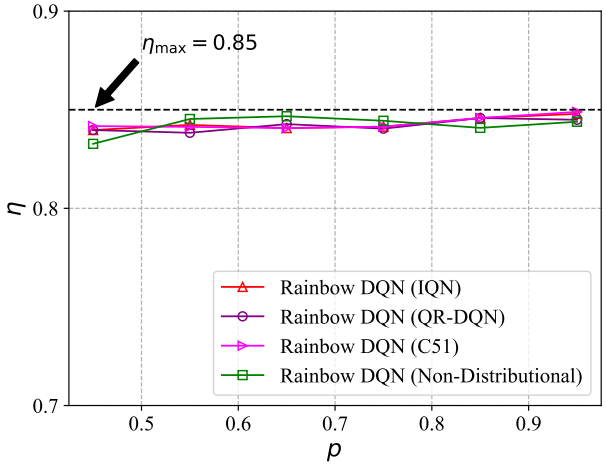

Fig. 4. The actual transmission cost (η) versus transmission successful probability p under different DRL algorithms. $\eta_{\max} = 0.85$ and $r_n = 0.7$.

tivity, we can define a risk-sensitive utility function

$$\mathcal{U}^{\pi}(s,a) = (1-\beta)Q^{\pi}(s,a) + \beta \mathrm{CVaR}_{\varphi}[Z^{\pi}(s,a)], \quad (14)$$

where $Q^{\pi}(s,a)$ is computed by

$$Q^{\pi}(s,a) = \frac{1}{K_{\text{sel}}} \sum_{k=1}^{K_{\text{sel}}} Z_{\tau_k}(s,a), \quad (15)$$

and $\beta \in [0,1]$ controls the trade-off between minimizing average VAoI and mitigating extreme VAoI. The scheduling policy using the risk-sensitive utility function is then given by

$$\pi^{\text{risk}}(s) = \arg\max_{a \in \mathcal{A}} \mathcal{U}^{\pi}(s,a). \quad (16)$$

By tuning β and φ, the scheduler can flexibly adopt risk-neutral ($\beta = 0$) or risk-averse ($\beta = 1$) strategies under different risk levels α. The risk-sensitive distributional DRL algorithm provides a robust and practical solution for controlling both average and extreme VAoI in multiuser scheduling under transmission constraints. Implementation based on IQN is summarized in Algorithm 1. Comparisons between different distributional DRL algorithms are given in the next section.

IV. Performance Evaluation

This section evaluates the proposed risk-sensitive distributional DRL scheduling strategy. We implement it within the Rainbow DQN framework [15], which integrates six improvements over DQN: Double DQN, Dueling DQN, Prioritized Replay, Multi-step Learning, Noisy Networks, and a distributional component (C51 by default). This setting allows direct comparison between different distributional variants and a non-distributional baseline, and also enables ablation studies to identify which Rainbow components contribute most to tail VAoI performance. While our focus is on tail risk, we also report average VAoI to examine potential trade-offs.

Simulation settings are as follows. We set the value range to $[-1500, 0]$ for C51, and $\mathcal{N} = \mathcal{N}' = 64$ quantiles for QR-DQN and IQN. All DRL algorithms consist of 50000 training episodes, with a learning rate of 0.002, a discount factor of $\gamma = 0.99$, and an ϵ-greedy exploration strategy with $\epsilon = 0.2$. The replay buffer has a size of 10^5 and the batch size is 128. The CVaR risk level is fixed at $\alpha = \varphi = 0.25$. Unless otherwise specified, we set $\beta = 0$ and $\beta = 1$ when the performance metric is the average VAoI and CVaR, respectively. The system has $N = 10$ users, and all users have the same packet generation rate of $r_n = 0.7$. For each setting of $\eta_{\max}$, p, and other parameters, we average results of five random seeds to ensure statistical reliability.

Fig. 3 presents the system-wide CVaR of VAoI across different distributional DRL algorithms (C51, QR-DQN, and IQN) and a non-distributional DQN baseline, all implemented within the Rainbow framework. We observe that all distributional approaches consistently achieve lower tail VAoI than the non-distributional baseline under different resource constraints $\eta_{\max}$. This is because by learning the full return distribution rather than just its expectation, distributional DRL explicitly models the likelihood and severity of tail events. Among the distributional DRL algorithms, IQN outperforms C51 and QR-DQN because its continuous quantile sampling allows flexible approximation of arbitrary quantiles, giving it superior ability to capture extreme VAoI outcomes and mitigate tail risks, which is consistent with the discussion in Section III.

Fig. 4 examines the actual transmission cost η under varying success probabilities p, with the maximum allowable cost $\eta_{\max} = 0.85$ indicated by the black dashed line. It shows that all algorithms remain below the constraint, validating the effectiveness of different algorithms. Importantly, distributional DRL achieves superior tail-risk mitigation as in Fig. 3 without violating resource budgets, demonstrating that these performance gains stem from more informed, distribution-aware scheduling rather than increased resource usage.

Next, we analyze IQN under different risk-tradeoff parameters β: (i) $\beta = 0$ (risk-neutral), (ii) $\beta = 1$ (risk-averse), and (iii) $\beta = 0.5$ (balanced). Both CVaR and average performance are considered. Fig. 5 shows that at $\eta_{\max} = 0.65$, the risk-averse setting ($\beta = 1$) reduces $\mathrm{CVaR}_{0.25}$ by up to 19.5% compared with the risk-neutral setting ($\beta = 0$), while the average VAoI increases slightly. In general, increasing

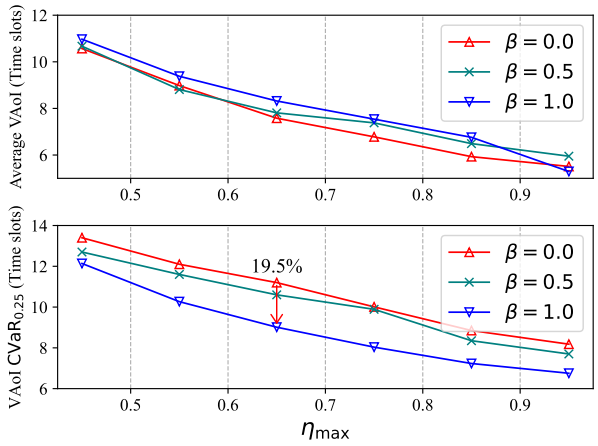

Fig. 5. The average VAoI (upper part) and the $\mathrm{CVaR}_{0.25}$ of system-wide VAoI (lower part) versus $\eta_{\max}$ when Rainbow with IQN is used. $p = 0.8$ and $r_n = 0.7$.

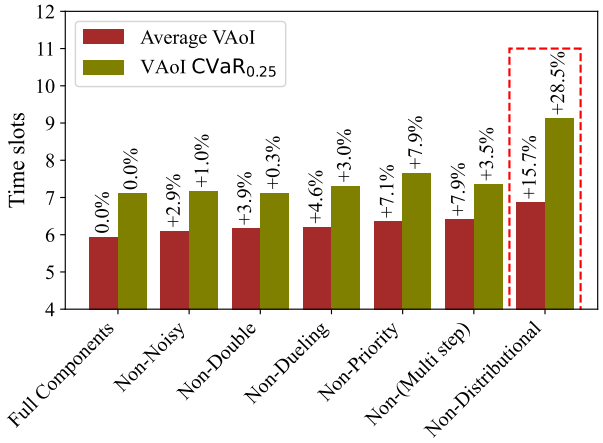

Fig. 6. Ablation Study of Rainbow with IQN on average VAoI and $\mathrm{CVaR}_{0.25}$ of system-wide VAoI under $p = 0.8$, $\eta_{\max} = 0.85$, and $r_n = 0.7$.

β systematically lowers the tail risk curve while slightly elevating the mean VAoI curve, illustrating a smooth and predictable trade-off. More specifically, risk-neutral agents prioritize average performance and occasionally incur extreme staleness, whereas risk-averse agents proactively avoid high-risk decisions by adopting more conservative scheduling. The parameter β therefore serves as a tunable and interpretable control knob, enabling designers to balance mean performance and reliability. It is critical for applications like industrial control systems where extreme latency must be avoided.

Finally, Fig. 6 reports an ablation study of Rainbow with IQN, evaluating the contribution of each module to both average VAoI and $\mathrm{CVaR}_{0.25}$ of VAoI. Simulations indicate that the ablation of other components produces smaller performance drops, while removing IQN causes a noticeable upward shift in average or CVaR of VAoI. Specifically, removing the IQN component results in the largest performance degradation, with average VAoI increasing by 15.7% and $\mathrm{CVaR}_{0.25}$ increasing by 28.5%. These results underscore the pivotal role of distributional modeling via IQN: it enables the agent to perceive and manage tail risks, fundamentally shifting decision-making from expectation-driven to risk-aware scheduling. By embedding risk sensitivity into the learning architecture, it provides a principled and practical framework for low-risk scheduling.

V. Conclusions

We have proposed a risk-sensitive scheduling strategy based on distributional DRL to address multiuser VAoI optimization under resource constraints. By formulating the problem as a CMDP and leveraging distributional modeling of returns, our approach captures extreme events, enabling optimization of the CVaR to manage tail risks. Numerical evaluations demonstrate that all distributional methods (C51, QR-DQN, IQN) outperform non-distributional baselines in reducing tail risk. IQN achieves more significant gains due to its flexible and accurate quantile approximation. Moreover, the proposed utility-based framework provides explicit control over the trade-off between mean performance and risk sensitivity, offering system designers an intuitive mechanism to tailor scheduling policies to reliability requirements. Overall, this work establishes distributional DRL with CVaR optimization as a principled, extensible, and practical solution for low-risk, high-freshness scheduling in real-time wireless systems.

References

[1] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *Proc. IEEE SECON*, Jun. 2011, pp. 350–358.

[2] R. D. Yates, "The age of gossip in networks," in *Proc. IEEE ISIT*, Jul. 2021, pp. 2984–2989.

[3] B. Buyukates, M. Bastopcu, and S. Ulukus, "Version age of information in clustered gossip networks," *IEEE J. Sel. Areas Inf. Theory*, vol. 3, no. 1, pp. 85–97, Mar. 2022.

[4] Y. Ji, Y. Lu, X. Xu, and X. Huang, "Age-optimal packet scheduling with resource constraint and feedback delay," *IEEE Trans. Commun.*, vol. 72, no. 7, pp. 4041–4054, Jul. 2024.

[5] Q. Chen et al., "Average AoI minimization with directional charging for wireless-powered network edge," *IEEE Trans. Mobile Comput.*, vol. 24, no. 6, pp. 4889–4906, Jun. 2025.

[6] E. Delfani and N. Pappas, "Semantics-aware status updates with energy harvesting devices: Query version age of information," in *Proc. WiOpt*, Oct. 2024, pp. 177–184.

[7] G. Karevvanavar, H. Pable, O. Patil, R. V. Bhat, and N. Pappas, "Version age of information minimization over fading broadcast channels," *IEEE Trans. Wireless Commun.*, vol. 24, no. 2, pp. 1620–1634, Feb. 2025.

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

[9] A. Avranas, P. Ciblat, and M. Kountouris, "Deep reinforcement learning for resource constrained multiclass scheduling in wireless networks," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 1, pp. 225–241, 2023.

[10] E. Eldeeb and H. Alves, "Offline and distributional reinforcement learning for wireless communications," *IEEE Commun. Mag.*, vol. 63, no. 8, pp. 71–76, Aug. 2025.

[11] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proc. ICML*, Aug. 2017, pp. 449–458.

[12] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *Proc. AAAI*, Feb. 2018, pp. 2892–2901.

[13] W. Dabney, G. Ostrovski, D. Silver, and R. Munos, "Implicit quantile networks for distributional reinforcement learning," in *Proc. ICML*, 2018, pp. 1096–1105.

[14] A. Kushwaha, K. Ravish, P. Lamba, and P. Kumar, "A survey of safe reinforcement learning and constrained MDPs: A technical survey on single-agent and multi-agent safety," 2025, *arXiv:2505.17342*.

[15] M. Hessel et al., "Rainbow: Combining improvements in deep reinforcement learning," in *Proc. AAAI*, 2018, pp. 3215–3222.