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

Uncertainty-Guided Adaptive Conservative Offline Reinforcement Learning for Safer Mechanical Ventilation

Huidong Liu[1], Hang Yu[2], Qiyang Zhang[1], Jiarui Dou[1], Xianlei Long[1], Jiantao Shi[3*], Fuqiang Gu[1*]
[1]College of Computer Science, Chongqing University
[2]School of Engineering, University of Warwick, Coventry, CV4 7AL, UK
[3]Department of Neurosurgery and Key Laboratory of Neurotrauma, Southwest Hospital, Third Military Medical University (Army Medical University)
{liuhuidong,zhangqiyang}@stu.cqu.edu.cn, hang.yu.1@warwick.ac.uk, { doujr, xianlei.long, gufq}@cqu.edu.cn, shijt@tmmu.edu.cn

Abstract

Mechanical ventilation (MV) is essential in intensive care units (ICUs), yet conventional protocols lack personalization and risk harmful over- or under-ventilation. Offline reinforcement learning (ORL) enables policy optimization from retrospective clinical data without unsafe online interaction, but existing methods are highly sensitive to distributional shift and out-of-distribution (OOD) actions, limiting their reliability in complex clinical settings. To address these challenges, we propose UBER-CQL (Uncertainty-Balanced Exploration and Robust Conservative Q-Learning), a robust ORL algorithm for safe decision-making under dataset shift. UBER-CQL integrates heteroscedastic Bayesian neural networks with conservative Q-learning to model posterior Q-value uncertainty, which is used to adaptively penalize unreliable high-risk actions while maintaining performance within the data support. We further design numerically stable objectives for conservative Bayesian value estimation. Experiments on in-distribution and OOD subsets of MIMIC-IV and eICU demonstrate that UBER-CQL outperforms state-of-the-art ORL and clinician baselines, producing safer and more effective MV strategies.

1 Introduction

Mechanical ventilation (MV) is a life-saving intervention widely used in intensive care units (ICUs) for patients with respiratory failure, including acute respiratory distress syndrome (ARDS) and chronic obstructive pulmonary disease (COPD) [Al-Husinat *et al.*, 2024]. Despite its clinical importance, determining optimal ventilator settings remains challenging due to the substantial heterogeneity across critically ill patients and the rapidly evolving nature of their physiological states. Current practice largely relies on clinician expertise or population-level protocols, which often lack the ability to adapt in a personalized and temporally consistent manner [Silva *et al.*, 2022; Umerenkov *et al.*, 2025]. These limitations motivate the need for data-driven approaches to individualized MV management.

Offline reinforcement learning (ORL) has recently emerged as a promising paradigm for learning treatment policies from retrospective electronic health record (EHR) data without unsafe online interaction [Yu *et al.*, 2021; Jayaraman *et al.*, 2024; Li *et al.*, 2025; Zhang *et al.*, 2025]. However, ORL in healthcare faces two critical challenges: (1) distributional shift, where learned policies select actions rarely observed in historical clinician data; and (2) Q-value overestimation for out-of-distribution (OOD) actions, which can lead to unsafe clinical recommendations.

Conservative ORL algorithms such as Conservative Q-Learning (CQL) [Yuan *et al.*, 2023] mitigate OOD overestimation by penalizing unseen actions, effectively constraining policies to the support of the dataset. While this improves safety, fixed conservative penalties often overly suppress useful policy improvement and lack flexibility across heterogeneous patient states. More recent uncertainty-aware approaches (e.g., ConformalDQN [Eghbali *et al.*, 2025]) incorporate calibrated risk control, but provide limited granularity in modeling epistemic uncertainty and its interaction with conservative value estimation.

To address these limitations, we propose UBER-CQL (Uncertainty-Balanced Exploration and Robust Conservative Q-Learning), a novel ORL framework for safe MV optimization (Figure 1). UBER-CQL introduces a heteroscedastic Bayesian neural network (HBNN) [Immer *et al.*, 2023] to

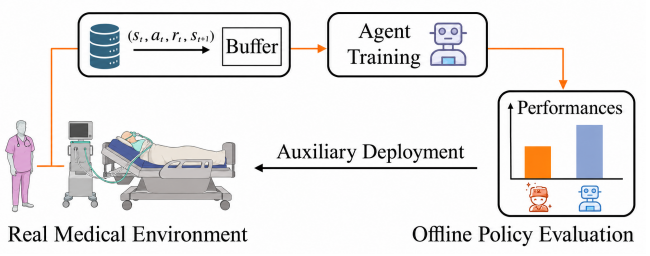


Figure 1: Illustration of ORL-based optimization for mechanical ventilation.

*Corresponding author

model the posterior distribution over Q-values, enabling fine-grained estimation of epistemic uncertainty. This uncertainty is incorporated into an adaptive penalty that dynamically adjusts conservatism according to state-specific confidence, allowing safer generalization beyond the behavior policy while avoiding unnecessary over-restriction. We further design numerically robust objectives to stabilize Bayesian conservative value learning. Then, we evaluate UBER-CQL on both MIMIC-IV and eICU cohorts, demonstrating consistent improvements over state-of-the-art ORL baselines and clinician policies across safety, policy value, and outcome-related metrics.

In summary, the main contributions of this work are as follows:

- We propose UBER-CQL, an uncertainty-aware conservative ORL framework that leverages HBNNs to model Q-value posteriors for reliable decision-making under distributional shift.
- We introduce an adaptive uncertainty-guided penalty that balances conservatism and policy improvement, reducing unsafe OOD actions while preserving performance.
- We develop numerically stable Bayesian conservative objectives and validate the framework on large-scale ICU datasets, where UBER-CQL consistently outperforms strong ORL and clinical baselines.

2 Related Work

Mechanical ventilation (MV) is a cornerstone therapy for critically ill patients with severe respiratory failure [Rubulotta *et al.*, 2024]. However, inappropriate ventilator settings can induce harmful complications, including ventilator-induced lung injury and diaphragm dysfunction [Powers, 2024]. Conventional MV management largely relies on standardized clinical guidelines or physician experience, which often fail to capture complex, patient-specific, and time-varying physiological dynamics [Liu *et al.*, 2024]. These limitations have motivated increasing interest in data-driven decision frameworks, particularly ORL, for personalized ventilator control.

VentAI [Peine *et al.*, 2021] represents one of the earliest RL-based systems for ventilator management, employing Q-learning to derive adaptive MV strategies from retrospective ICU data. This work demonstrated the feasibility of sequential decision-making for personalized ventilation. Subsequently, Yin et al. [Yin *et al.*, 2022] proposed the Deconfounding Actor-Critic (DAC) network, which integrates short- and long-term reward modeling with deconfounding mechanisms to mitigate bias and improve treatment policy accuracy. To address extrapolation errors under offline settings, DeepVent [Kondrup *et al.*, 2023] adopted CQL with intermediate reward shaping to constrain learned policies within the empirical data support. While conservative methods improve safety by discouraging unsupported actions, their fixed penalty design often limits policy flexibility and reduces adaptability across heterogeneous patient states. ConformalDQN [Eghbali *et al.*, 2025] introduced conformal prediction into deep Q-learning to quantify uncertainty and detect OOD actions. This approach provides distribution-free calibration and improves robustness; however, uncertainty is incorporated only post hoc, preventing dynamic regulation of conservative penalties during training and limiting fine-grained action-level risk control [Yue *et al.*, 2023; Zhang *et al.*, 2024].

Despite recent progress, two key limitations persist: (1) most ORL approaches rely on static conservatism that cannot adapt to varying uncertainty across patient states; and (2) existing methods lack precise state–action uncertainty modeling to guide risk-sensitive policy learning in safety-critical clinical environments. To overcome these challenges, we propose UBER-CQL, a unified ORL framework that integrates HBNNs with conservative Q-learning. By jointly modeling Q-value means and variances, UBER-CQL enables fine-grained epistemic uncertainty estimation and adaptively penalizes high-risk actions during training, resulting in improved robustness, safer OOD handling, and more personalized MV decision-making.

3 Problem Formulation

Personalized MV involves dynamically adjusting ventilator parameters over the course of treatment to maintain physiological stability—e.g., optimizing blood oxygen saturation and arterial carbon dioxide levels—while reducing risks such as ventilator-induced lung injury. The ultimate goal is to improve patient outcomes and survival.

We formulate MV management as a continuous-time control problem. The patient's physiological state at time t is represented as $x(t) \in \mathbb{R}^n$, and the ventilator control variables (e.g., tidal volume, PEEP) as $u(t) \in \mathbb{R}^m$. The state evolves according to nonlinear dynamics:

$$\dot{x}(t) = f(x(t), u(t)), \tag{1}$$

where $f(\cdot)$ denotes unknown patient-specific transition dynamics. The objective is to determine a control policy $u(t)$ that minimizes the cumulative treatment cost:

$$J = \int_{t_0}^{t_f} L(x(t), u(t))\, dt + \Phi(x(t_f)), \tag{2}$$

where $L(\cdot)$ captures instantaneous clinical risk and physiological deviation, and $\Phi(\cdot)$ represents the terminal outcome-related cost.

Since the transition function $f(\cdot)$ is unknown and online interaction is infeasible in safety-critical clinical environments, we reformulate personalized MV management as an ORL problem under a Markov Decision Process (MDP) framework [Ernst and Louette, 2024]. The MDP is defined as $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$, where:

- $\mathcal{S}$: state space representing patient physiology. We construct a 42-dimensional state vector comprising demographic information, vital signs, laboratory measurements, and fluid-related indicators, collectively capturing the patient's clinical status.
- $\mathcal{A}$: action space corresponding to ventilator settings. Actions consist of three controllable parameters: tidal volume adjusted by ideal body weight (Vt), positive

end-expiratory pressure (PEEP), and fraction of inspired oxygen (FiO_2). Each parameter is discretized into seven levels, yielding a combinatorial action space:$\mathcal{A} = Vt \times PEEP \times FiO_2$.

- $P(s'|s,a)$: transition probability from state s to s' under action a. As direct interaction with patients is impossible, transitions are implicitly learned from the offline dataset $\mathcal{D} = \{(s,a,r,s')\}$, enabling value estimation based on empirical clinical dynamics.
- $r(s,a)$: reward function measuring treatment effectiveness. We define rewards using both short-term physiological improvement and long-term survival outcomes. Intermediate rewards are derived from normalized changes in the Apache II score, while terminal rewards encode 90-day mortality:

$$r(s,a) = \begin{cases} +1, & \text{if the patient survives,} \\ -1, & \text{if the patient dies,} \\ \dfrac{A_t - A_{t+1}}{A_{\max} - A_{\min}}, & \text{otherwise,} \end{cases} \tag{3}$$

  where A_t denotes the Apache II score at time t, and $A_{\max}, A_{\min}$ are used for normalization. This design ensures alignment between immediate physiological stabilization and long-term prognosis [Eghbali *et al.*, 2025].
- $\gamma \in (0,1)$: discount factor controlling the relative importance of immediate versus future rewards. Smaller values emphasize short-term physiological stabilization, while values closer to 1 prioritize long-term clinical outcomes.

Given an offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$, ORL aims to learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected discounted return:

$$J(\pi) = \mathbb{E}_{\xi\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right], \tag{4}$$

where $\xi = (s_0, a_0, s_1, a_1, \dots)$ denotes a trajectory induced by policy π.

4 Proposed Method: UBER-CQL

In this section, we present the proposed UBER-CQL framework for personalized mechanical ventilation optimization. As illustrated in Figure 2, the framework consists of three stages: 1) *Data Extraction*, which preprocesses 42 physiological variables to construct an offline MDP dataset; 2) *UBER-CQL*, an uncertainty-aware conservative RL algorithm for safe ventilator control; and 3) *Evaluation*, where offline policy evaluation estimates policy value on held-out cohorts. We detail the UBER-CQL algorithm below and describe Data Extraction and Evaluation in the following section.

4.1 Uncertainty-Aware Q-Value Estimation via HBNNs

Offline RL for MV optimization faces substantial challenges due to limited data coverage and the prevalence of OOD actions. These issues can result in unsafe control decisions, as the learned policy may query state-action pairs that are underrepresented in the dataset.

To address this, we adopt a HBNN to model uncertainty in Q-value estimation. Unlike standard Bayesian neural networks that assume homoscedastic noise, HBNNs capture input-dependent predictive variance, allowing for the estimation of both the predictive mean and aleatoric uncertainty, which reflects irreducible variability in clinical data [Skafte *et al.*, 2019].

Formally, let θ denote the parameters of the HBNN, which models the Q-value for each state-action pair (s,a) as a Gaussian distribution:

$$Q_\theta(s,a) \sim \mathcal{N}\left(\mu_\theta(s,a), \sigma_\theta^2(s,a)\right), \tag{5}$$

where both the predictive mean $\mu_\theta : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ and variance $\sigma_\theta^2 : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^+$ are learned. This probabilistic formulation contrasts with deterministic Q-networks $Q_\phi(s,a)$, which yield only point estimates.

Training proceeds by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{HBNN}} = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\frac{(y-\mu_\theta(s,a))^2}{2\sigma_\theta^2(s,a)} + \frac{1}{2}\log\sigma_\theta^2(s,a)\right], \tag{6}$$

where the target value $y = r + \gamma \max_{a'} \mu_{\theta^-}(s',a')$ is computed using a target network θ^-. This loss formulation naturally discourages high variance in data-scarce regions, encouraging conservative and risk-aware action selection—critical for ensuring safety in high-stakes clinical settings like MV therapy.

4.2 Adaptive Uncertainty Management and Action Penalization

To further enhance robustness against OOD actions and distribution shift in offline MV datasets, we propose a dynamic uncertainty-aware learning framework. By leveraging model uncertainty to modulate Q-value targets, the framework mitigates overestimation bias while preventing overly pessimistic updates. As a result, it enables safer and more personalized policy optimization tailored for high-stakes MV settings, where decision reliability is paramount. This framework synergistically combines three key components: uncertainty-guided action penalization, adaptive Bellman updates, and conservative Q-learning regularization.

Uncertainty-Guided Action Penalization

In safety-critical MV tasks, actions exhibiting high epistemic uncertainty—stemming from sparse data coverage in offline datasets—can precipitate hazardous clinical outcomes, such as barotrauma induced by excessive airway pressures or volutrauma from overdistension. To counteract this, we utilize a deep ensemble of $K = 5$ HBNNs for robust uncertainty quantification [Seitzer *et al.*, 2022], where the ensemble mean $\bar{Q}_\theta(s,a) = \frac{1}{K}\sum_{k=1}^K \mu_{\theta_k}(s,a)$ and epistemic uncertainty $\sigma_\theta^{\text{ep}}(s,a) = \text{std}_k\{\mu_{\theta_k}(s,a)\}$ are computed across members. Each individual HBNN produces a predictive mean Q-value and an output variance that models aleatoric uncertainty (i.e., inherent stochasticity in observations, such as variable patient responses [Simone *et al.*, 2026]), while epistemic uncertainty reflects model ignorance in underrepresented regions. This

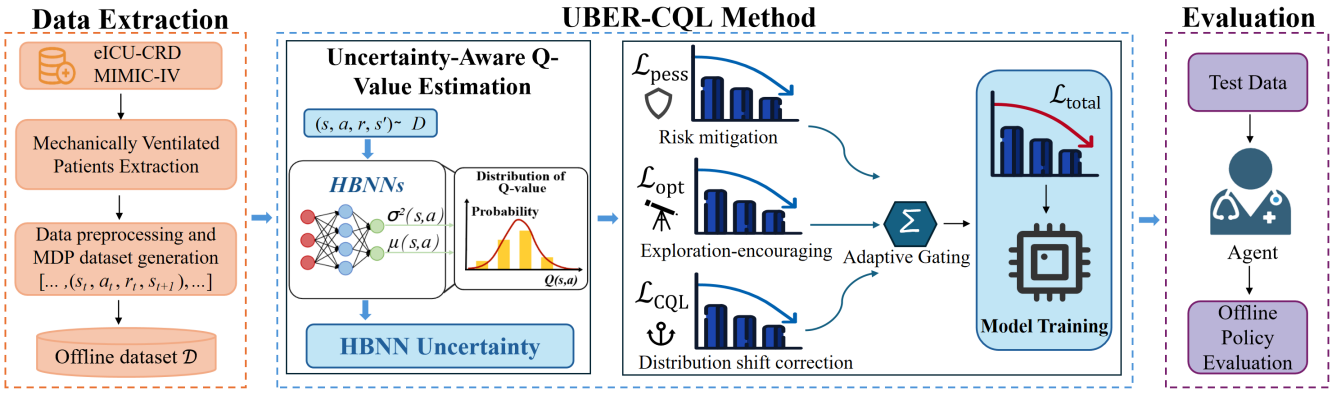


Figure 2: Overall pipeline of the proposed UBER-CQL framework for uncertainty-aware offline optimization of mechanical ventilation.

ensemble-based approach serves as an efficient approximation to the posterior distribution over network parameters, enabling reliable inference without the computational overhead of full Bayesian methods.

To penalize high-uncertainty actions, we compute a dynamic threshold τ_σ as the p_u-th percentile of $\{\sigma_\theta^{\text{ep}}(s,a) \mid (s,a) \in \mathcal{B}\}$, where $\mathcal{B}$ is a mini-batch from dataset $\mathcal{D}$ and $p_u = 0.8$:

$$\tau_\sigma = \text{Quantile}\left(\{\sigma_\theta^{\text{ep}}(s,a) \mid (s,a) \in \mathcal{B}\}, p_u\right). \tag{7}$$

The penalized Q-function is then:

$$Q_{\text{p}}(s,a) = \begin{cases} \bar{Q}_\theta(s,a) - \beta \cdot \sigma_\theta^{\text{ep}}(s,a), & \text{if } \sigma_\theta^{\text{ep}}(s,a) > \tau_\sigma, \\ \bar{Q}_\theta(s,a), & \text{otherwise,} \end{cases} \tag{8}$$

where $\bar{Q}_\theta(s,a)$ is the ensemble mean and $\beta = 0.5$.

This selective penalization mechanism reduces the agent's tendency to exploit poorly supported regions of the state-action space, such as extreme ventilation parameters (e.g., abnormally high tidal volumes or PEEP levels), thereby lowering the risk of adverse outcomes such as barotrauma or volutrauma. By incorporating action-specific epistemic uncertainty into the value estimation pipeline, this approach provides a principled mechanism to encourage cautious decision-making under uncertainty—an essential property for real-world clinical deployment.

Adaptive Bellman Updates

To balance exploration and conservatism in offline reinforcement learning, we introduce an adaptive Bellman update mechanism that interpolates between optimistic and pessimistic targets. This dynamic weighting enables the agent to benefit from exploratory value propagation in well-supported regions while remaining conservative under high uncertainty or distribution shift.

The *optimistic* Bellman loss encourages standard value iteration by following the classical Bellman optimality operator:

$$\mathcal{L}_{\text{opt}} = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(\mu_\theta(s,a) - (r + \gamma \max_{a'} \mu_{\theta^-}(s',a'))\right)^2\right], \tag{9}$$

where r is the observed reward, γ is the discount factor, and μ_{θ^-} denotes the target network updated with a delay to stabilize learning.

In contrast, the *pessimistic* Bellman loss incorporates uncertainty-aware penalization to discourage overestimation in regions with limited support. It replaces the Q-value target with the penalized value $Q_{\text{p}}(s',a')$ defined earlier:

$$\mathcal{L}_{\text{pess}} = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(\mu_\theta(s,a) - (r + \gamma \max_{a'} Q_{\text{p}}(s',a'))\right)^2\right]. \tag{10}$$

The adaptive Bellman update mechanism is formalized by combining the optimistic and pessimistic losses into a unified adaptive loss function, weighted dynamically based on uncertainty or policy confidence:

$$\mathcal{L}_{\text{adaptive}} = \omega \mathcal{L}_{\text{opt}} + (1-\omega)\mathcal{L}_{\text{pess}}, \tag{11}$$

where ω dynamically adjusts the trade-off between optimistic and pessimistic Bellman updates. Specifically, ω decays over training to gradually shift focus from exploration to conservatism:

$$\omega = (\omega_{\max} - \omega_{\min}) \cdot \exp(-3 \cdot p) + \omega_{\min}, \tag{12}$$

where $\omega_{\max} = 0.9$, $\omega_{\min} = 0.1$, and $p = \min(\text{iterations}/30000, 1)$ represents normalized training progress. This schedule ensures the policy is exploratory early on and progressively conservative for safe deployment in clinical settings.

Conservative Q-Learning Regularization

To further reduce overestimation bias prevalent in offline RL, we incorporate a CQL penalty that discourages assigning high Q-values to unsupported actions:

$$\mathcal{L}_{\text{CQL}} = \alpha \cdot \left(\mathbb{E}_{s\sim\mathcal{D}}\left[\log \sum_a \exp\left(\mu_\theta(s,a)\right)\right] - \mathbb{E}_{(s,a)\sim\mathcal{D}}[\mu_\theta(s,a)]\right), \tag{13}$$

where $\alpha = 0.1$ is the regularization coefficient. This term penalizes overconfident Q-values for out-of-distribution actions, enhancing safety and robustness under limited data coverage.

Statistic	MIMIC-IV	eICU-CRD
Cohort Demographics		
Number of Patients (Episodes)	37,615	28,562
Age (years, Mean)	64.59	63.05
Hospital LOS (hours, Mean)	282.58	236.18
Mortality (Count, %)	9,369 (24.9%)	9,795 (20.0%)
MDP Statistics		
Total Transitions (Steps)	319,837	116,591
Avg. Episode Length (Steps)	8.50	4.08
Median Episode Length	6.00	2.00
Avg. Episode Return (Raw)	0.50	0.58

Table 1: Statistics of processed MIMIC-IV and eICU.

4.3 Training

The overall training objective integrates all components to guide uncertainty-aware and conservative policy learning:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adaptive}} + \lambda \mathcal{L}_{\text{HBNN}} + \mathcal{L}_{\text{CQL}}, \tag{14}$$

where $\lambda = 0.1$ balances the heteroscedastic Bayesian Q-value loss.

5 Experiments and Results

5.1 Experimental Setup

We extracted mechanically ventilated patient cohorts from two large-scale ICU databases: MIMIC-IV [Johnson *et al.*, 2023] and eICU [Pollard *et al.*, 2018]. To rigorously evaluate cross-database generalization, we use MIMIC-IV as the primary training source and eICU as an independent cross-database test cohort. Demographic and physiological statistics for the resulting Markov Decision Processes (MDPs) are summarized in Table 1.

Following prior work [Eghbali *et al.*, 2025], we also curated an out-of-distribution (OOD) cohort representing physiologically atypical patients. A patient is labeled as OOD if any initial state feature—including demographics, vital signs, laboratory results, or fluid status—falls within the top or bottom 1% of the population distribution. This procedure yields an OOD subset comprising approximately 25% of the cohort. Both in-distribution (ID) and OOD datasets were split into training (60%), validation (20%), and testing (20%) sets. Training and validation data were used for model fitting and hyperparameter tuning, while the held-out test set was reserved for final evaluation.

5.2 Evaluation

To assess the performance of learned policies in our offline RL framework for mechanical ventilation, we employ two established off-policy evaluation (OPE) methods: Fitted Q-Evaluation (FQE) [Hao *et al.*, 2021] and the Doubly Robust (DR) estimator [Tang and Wiens, 2021]. These approaches enable reliable estimation of expected returns using historical data without requiring online interaction, which is critical in clinical settings.

FQE approximates the optimal Q-function Q^* by minimizing the mean squared Bellman error over the offline dataset:

$$Q^{\pi} = \arg\min_{Q \in \mathcal{F}} \frac{1}{M} \sum_{i=1}^{M} \left(Q(s_i, a_i) - Q_{\mathrm{t},i}\right)^2, \tag{15}$$

where $\mathcal{F}$ denotes the Q-function hypothesis space, M is the number of samples, and $Q_{\mathrm{t},i}$ is the Bellman target computed using the evaluation policy π.

Let $\delta_t^{(i)} = r_t^{(i)} + \gamma \hat{V}(s_{t+1}^{(i)}) - \hat{Q}(s_t^{(i)}, a_t^{(i)})$ denote the single-step Bellman residual under policy π. The step-wise DR estimator is then:

$$\hat{V}_{\text{DR}}^{\text{step}} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{V}(s_0^{(i)}) + \sum_{t=0}^{H_i - 1} \gamma^t \, w'_{t,i} \, \delta_t^{(i)} \right], \tag{16}$$

where $w'_{t,i} = \min\left(\prod_{k=0}^{t} \frac{\pi(a_k^{(i)} | s_k^{(i)})}{\mu(a_k^{(i)} | s_k^{(i)})}, C \right)$ is the clipped cumulative importance weight, and $C > 0$ is the clipping threshold.

5.3 Baselines and Implementation Details

To evaluate the effectiveness of UBER-CQL for mechanical ventilation optimization, we compare it against several established offline RL baselines relevant to clinical decision-making: DDQN [Feng *et al.*, 2023], BCQ [Liu *et al.*, 2024], CQL [Kondrup *et al.*, 2023], and ConformalDQN [Eghbali *et al.*, 2025].

All models were implemented in Python 3.8 using PyTorch 1.13. We adopt a shared architecture with two hidden layers of 256 units and ReLU activations. Training proceeded for up to 30,000 timesteps, with target networks updated every 5,000 steps for stability. Optimal hyperparameters, including learning rates $[10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$ and discount factors $\gamma \in [0.05, 0.1, 0.5, 0.75, 0.9]$, were selected via grid search on the validation set, yielding $\gamma = 0.75$. Experiments were performed on Ubuntu 20.04.5 LTS with an Intel Core i7-10700 CPU, NVIDIA RTX 3090 Ti GPU, and 32 GB RAM. All models were trained over five independent runs, and reported results correspond to the mean performance to ensure robustness and reproducibility.

For UBER-CQL, the best performance was achieved with a learning rate of 5×10^{-4}. Baseline hyperparameters were configured as follows: BCQ (lr $= 1 \times 10^{-4}$); CQL (lr $= 1 \times 10^{-4}$); ConformalDQN (lr $= 1 \times 10^{-3}$, $\alpha = 0.15$); and DDQN (lr $= 1 \times 10^{-3}$).

5.4 Results and Analysis

Quantitative Results

We evaluate model effectiveness in optimizing MV treatment by examining the Pearson correlation coefficient [Checkley *et al.*, 2008] between predicted Q-values and patient mortality. For each initial state-action pair (s, a) in the test set, we compute the predicted Q-value $Q(s, a)$ and the corresponding binary mortality outcome $M(s, a)$ (1 for death, 0 for survival).

As reported in Table 2, all models exhibit negative correlations, indicating that higher predicted Q-values are associated with lower mortality. The proposed UBER-CQL achieves the strongest correlations of -0.538 ± 0.007 on MIMIC-IV and -0.414 ± 0.009 on eICU, outperforming the strongest baseline (ConformalDQN) by approximately 19% due to its uncertainty-aware regularization. These results demonstrate the superior robustness and clinical reliability of UBER-CQL across heterogeneous ICU datasets.

Prior work [Luo *et al.*, 2024] has highlighted limitations in existing RL methods regarding evaluation consistency and baseline selection, which can lead to misleading performance claims. To contextualize this, we include simple baseline policies such as random, zero-drug, and max-drug strategies. As shown in Table 3, these naïve baselines achieve near-zero initial FQE scores on both MIMIC-IV and eICU, reflecting their limited ability to anticipate long-term outcomes from the onset of mechanical ventilation. Among offline RL methods, CQL outperforms BCQ, while the uncertainty-aware ConformalDQN further improves policy value estimation. The proposed UBER-CQL achieves higher estimated value estimates across both datasets, with FQE scores of $\mathbf{0.815 \pm 0.021}$ on MIMIC-IV and $\mathbf{0.725 \pm 0.022}$ on eICU, exceeding standard baselines, uncertainty-aware methods, and even the physician policy. These results demonstrate that incorporating fine-grained uncertainty modeling within conservative Q-learning produces more reliable and robust value estimates in offline clinical settings.

Robustness Evaluation

To evaluate robustness under distribution shifts, we compare mean initial Q-values for ID and OOD states, providing insight into each model's susceptibility to Q-value overestimation—a critical factor for safe and reliable decision-making in MV.

Figure 3 presents the mean estimated Q-values for ID and OOD states, with the red dashed line indicating the overestimation threshold at $Q = 1.0$. Traditional methods such as DDQN exhibit substantial overestimation, particularly in OOD settings, with means of 1.112 (ID) and 1.237 (OOD), both exceeding the threshold. CQL yields more conservative estimates, with means of 0.789 (ID) and 0.840 (OOD), remaining below 1.0. BCQ produces consistently low and stable values (0.710 ID, 0.762 OOD), suggesting slight underestimation. ConformalDQN, leveraging distribution-free uncertainty quantification, achieves reasonable calibration with 0.728 (ID) and 0.777 (OOD), mitigating overestimation while retaining responsiveness. Notably, UBER-CQL achieves the most balanced and stable estimates, with 0.899 (ID) and 0.894 (OOD), staying well below the threshold and showing minimal degradation under distribution shifts. In OOD scenarios, UBER-CQL's Q-values are approximately 27.8% lower than DDQN's and 15.1% higher than ConformalDQN's, demonstrating its effective trade-off between avoiding overestimation and preventing undue conservatism.

These results highlight the efficacy of UBER-CQL's uncertainty-aware penalization in managing distributional shifts and promoting cautious decision-making under high epistemic uncertainty, ensuring reliable value estimation for high-stakes ICU interventions.

Model	Correlation Coefficient (Mean ± Std)	
	MIMIC-IV	eICU
DDQN [Feng *et al.*, 2023]	-0.347 ± 0.029	-0.225 ± 0.033
CQL [Kondrup *et al.*, 2023]	-0.435 ± 0.032	-0.293 ± 0.029
BCQ [Liu *et al.*, 2024]	-0.424 ± 0.044	-0.316 ± 0.047
ConformalDQN [Eghbali *et al.*, 2025]	-0.453 ± 0.011	-0.359 ± 0.017
UBER-CQL	**-0.538 ± 0.007**	**-0.414 ± 0.009**

Table 2: Pearson correlation coefficients between predicted Q-values and patient mortality across different policy models on MIMIC-IV and eICU datasets. Results represent mean ± standard deviation over five independent runs. Higher-magnitude negative correlations indicate stronger alignment between predicted Q-values and improved patient outcomes.

Model	MIMIC-IV		eICU	
	FQE	DR	FQE	DR
Physician	0.478±0.003	-	0.560±0.009	-
Random Policy	0.013±0.006	-	0.014±0.002	-
Zero-Drug Policy	0.011±0.141	-	0.014±0.137	-
Max-Drug Policy	0.016±0.056	-	0.014±0.051	-
CQL	0.532±0.041	2.185 ± 0.035	0.438±0.001	1.014 ± 0.054
BCQ	0.352±0.001	0.842 ± 0.025	0.331 ± 0.001	0.418 ± 0.066
ConformalDQN	0.711 ± 0.003	4.042 ± 0.072	0.669 ± 0.003	1.922 ± 0.114
UBER-CQL	**0.815 ± 0.021**	**4.292 ± 0.083**	**0.725 ±0.022**	**2.179 ± 0.16**

Table 3: Mean initial FQE and step-wise DR estimates across different policy models on MIMIC-IV and eICU datasets. Results are reported as mean ± standard deviation over five independent runs. Higher values indicate better expected policy performance.

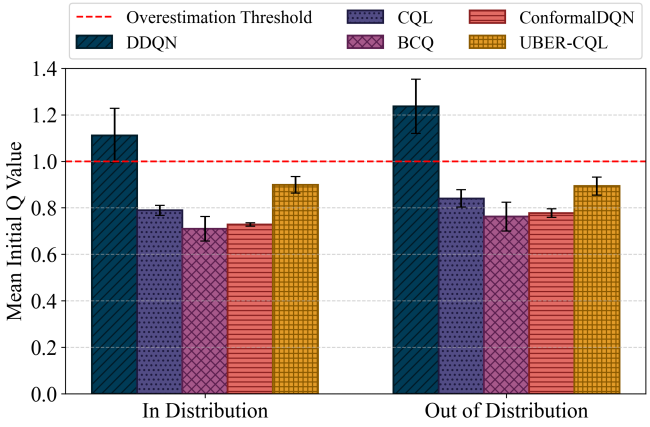


Figure 3: Mean estimated initial Q-values for in-distribution and out-of-distribution states. The red dashed line indicates the overestimation threshold at (Q=1.0).

Method	MIMIC-IV		eICU	
	FQE	DR	FQE	DR
w/o Opt/Pess	0.741±0.032	3.692±0.062	0.615±0.016	1.892±0.035
w/o Adaptive Bellman	0.781 ± 0.003	4.196 ± 0.071	0.698 ± 0.003	2.065 ± 0.115
w/o CQL	0.587±0.003	2.164±0.037	0.523±0.003	1.117±0.088
UBER-CQL	**0.815 ± 0.021**	**4.292 ± 0.083**	**0.725 ±0.022**	**2.179 ± 0.16**

Table 4: Ablation study assessing the contribution of each component in UBER-CQL. Removing the optimistic/pessimistic loss, adaptive Bellman updates, or CQL regularization reduces policy performance, demonstrating the importance of each module for robust offline MV optimization.

Ablation Study

We conducted ablation experiments to systematically assess the contribution of each core component in the UBER-CQL framework.

Impact of Bellman Update Strategies. Table 4 reports the ablation results on MIMIC-IV and eICU datasets under both FQE and DR evaluation. The full UBER-CQL consistently achieves the highest performance across all metrics, confirming the effectiveness of our uncertainty-aware and conservative learning design.

First, we examine the role of the Bellman update strategy. The variant *w/o Opt/Pess*, which removes both optimistic and pessimistic Bellman objectives and relies on a standard Bellman update, exhibits clear performance degradation on both datasets. This highlights that explicitly incorporating optimistic and pessimistic value targets is crucial for robust Q-value estimation in offline RL.

Next, disabling the adaptive interpolation between optimistic and pessimistic updates (*w/o Adaptive Bellman*) also reduces performance relative to the full model. This indicates that not only are the dual objectives important, but dynamically balancing them is essential for effectively reconciling exploration with conservatism under distributional shift.

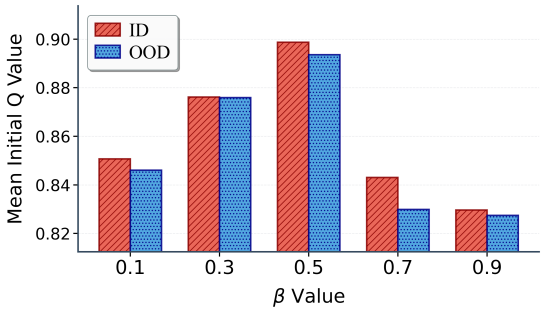
Finally, we evaluate the contribution of conservative Q-learning regularization. Removing the CQL penalty (*w/o CQL*) leads to the largest performance drop, particularly in FQE scores, underscoring its critical role in mitigating overestimation and improving robustness against OOD actions. The magnitude of this drop emphasizes that CQL complements the adaptive Bellman updates to stabilize policy learning.

Overall, these results demonstrate that each component—optimistic/pessimistic Bellman targets, adaptive interpolation, and conservative Q-learning—contributes meaningfully to UBER-CQL's performance. Their joint integration is key to achieving reliable, safe, and effective offline policy learning for high-stakes mechanical ventilation control.

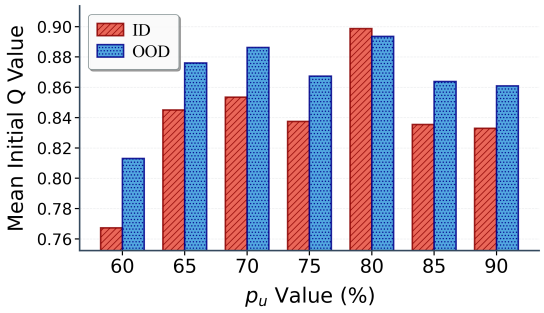
Sensitivity Analysis of Key Hyperparameters. We evaluate the sensitivity of UBER-CQL to two key hyperparameters: the uncertainty penalization coefficient β and the uncertainty quantile threshold p_u, under both ID and OOD settings (Figure 4).

As shown in Figure 4(a), increasing β from 0.1 to 0.5 improves the mean initial Q-values for both ID and OOD actions, suggesting that a moderate penalty on high-uncertainty actions effectively suppresses overestimation in under-supported regions. However, larger values of β lead to performance degradation, particularly for OOD actions, as overly aggressive penalization may incorrectly suppress Q-values for actions that are clinically valid but epistemically uncertain due to limited data coverage.

Figure 4(b) shows that increasing p_u from 60% to 80% enhances performance by better penalizing uncertain actions, but further increases yield diminishing returns, potentially restricting optimal ventilation strategies in ICU scenarios. Overall, these findings demonstrate UBER-CQL's robustness across reasonable hyperparameter ranges and offer practical tuning guidance.



(a) Sensitivity to the uncertainty penalization coefficient β.



(b) Sensitivity to the uncertainty quantile threshold p_u.

Figure 4: Sensitivity analysis of key hyperparameters in UBER-CQL.The results illustrate how varying these hyperparameters affects policy value estimates, highlighting the robustness of the proposed framework across reasonable parameter ranges.

6 Conclusion

We proposed UBER-CQL, a novel ORL framework tailored for safety-critical clinical decision-making, with a focus on personalized MV. UBER-CQL unifies HBNNs with conservative Q-learning to jointly model action-value uncertainty and adaptive conservatism. By explicitly estimating both the mean and variance of Q-values, the method provides fine-grained epistemic uncertainty that is leveraged to dynamically penalize high-risk and poorly supported actions, thereby improving robustness under distribution shift and mitigating extrapolation errors common in offline settings. Extensive experiments on real-world MV cohorts from MIMIC-IV and eICU demonstrate that UBER-CQL consistently outperforms state-of-the-art offline RL methods and clinical baselines in terms of value estimation, robustness to OOD states, and alignment with patient survival outcomes. These results highlight the promise of uncertainty-aware conservative learning for safe and reliable policy optimization in high-stakes healthcare applications. Future work will investigate extensions to multi-modal clinical data, as well as the integration of causal inference to further improve interpretability and decision reliability across broader clinical domains.

Acknowledgments

This paper is supported by the National Key Research and Development Program for Young Scientists (No. 2025YFB4711500), National Natural Science Foundation of

China (No. 42474027, No. 62403085), Xiaomi Young Talents Program and Graduate Research and Innovation Foundation of Chongqing, China (No. CYB240066).

References

[Al-Husinat *et al.*, 2024] Lou'i Al-Husinat, Saif Azzam, Sarah Al Sharie, Ahmed H Al Sharie, Denise Battaglini, Chiara Robba, John J Marini, Lauren T Thornton, Fernanda F Cruz, Pedro L Silva, et al. Effects of mechanical ventilation on the interstitial extracellular matrix in healthy lungs and lungs affected by acute respiratory distress syndrome: a narrative review. *Critical Care*, 28(1):165, 2024.

[Checkley *et al.*, 2008] William Checkley, Roy Brower, Anna Korpak, and B Taylor Thompson. Effects of a clinical trial on mechanical ventilation practices in patients with acute lung injury. *American journal of respiratory and critical care medicine*, 177(11):1215–1222, 2008.

[Eghbali *et al.*, 2025] Niloufar Eghbali, Tuka Alhanai, and Mohammad M Ghassemi. Distribution-free uncertainty quantification in mechanical ventilation treatment: A conformal deep q-learning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27960–27968, 2025.

[Ernst and Louette, 2024] Damien Ernst and Arthur Louette. Introduction to reinforcement learning. *Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P*, pages 111–126, 2024.

[Feng *et al.*, 2023] Xue Feng, Daoyuan Wang, Qing Pan, Molei Yan, Xiaoqing Liu, Yanfei Shen, Luping Fang, Guolong Cai, and Gangmin Ning. Reinforcement learning model for managing noninvasive ventilation switching policy. *IEEE Journal of Biomedical and Health Informatics*, 27(8):4120–4130, 2023.

[Hao *et al.*, 2021] Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR, 2021.

[Immer *et al.*, 2023] Alexander Immer, Emanuele Palumbo, Alexander Marx, and Julia Vogt. Effective bayesian heteroscedastic regression with deep neural networks. *Advances in Neural Information Processing Systems*, 36:53996–54019, 2023.

[Jayaraman *et al.*, 2024] Pushkala Jayaraman, Jacob Desman, Moein Sabounchi, Girish N Nadkarni, and Ankit Sakhuja. A primer on reinforcement learning in medicine for clinicians. *NPJ Digital Medicine*, 7(1):337, 2024.

[Johnson *et al.*, 2023] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[Kondrup *et al.*, 2023] Flemming Kondrup, Thomas Jiralerspong, Elaine Lau, Nathan de Lara, Jacob Shkrob, My Duc Tran, Doina Precup, and Sumana Basu. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15696–15702, 2023.

[Li *et al.*, 2025] Yikuan Li, Chengsheng Mao, Kaixuan Huang, Hanyin Wang, Zheng Yu, Mengdi Wang, and Yuan Luo. Deep reinforcement learning for efficient and fair allocation of health care resources. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 9790–9798, 2025.

[Liu *et al.*, 2024] Siqi Liu, Qianyi Xu, Zhuoyang Xu, Zhuo Liu, Xingzhi Sun, Guotong Xie, Mengling Feng, and Kay Choong See. Reinforcement learning to optimize ventilator settings for patients on invasive mechanical ventilation: retrospective study. *Journal of Medical Internet Research*, 26:e44494, 2024.

[Luo *et al.*, 2024] Zhiyao Luo, Yangchen Pan, Peter Watkinson, and Tingting Zhu. Reinforcement learning in dynamic treatment regimes needs critical reexamination. *arXiv preprint arXiv:2405.18556*, 2024.

[Peine *et al.*, 2021] Arne Peine, Ahmed Hallawa, Johannes Bickenbach, Guido Dartmann, Lejla Begic Fazlic, Anke Schmeink, Gerd Ascheid, Christoph Thiemermann, Andreas Schuppert, Ryan Kindle, et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *NPJ digital medicine*, 4(1):32, 2021.

[Pollard *et al.*, 2018] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

[Powers, 2024] Scott K Powers. Ventilator-induced diaphragm dysfunction: phenomenology and mechanism (s) of pathogenesis. *The Journal of Physiology*, 602(19):4729–4752, 2024.

[Rubulotta *et al.*, 2024] Francesca Rubulotta, Lluis Blanch Torra, Kuban D Naidoo, Hatem Soliman Aboumarie, Lufuno R Mathivha, Abdulrahman Y Asiri, Leonardo Sarlabous Uranga, and Sabri Soussi. Mechanical ventilation, past, present, and future. *Anesthesia & Analgesia*, 138(2):308–325, 2024.

[Seitzer *et al.*, 2022] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022.

[Silva *et al.*, 2022] PL Silva, PRM Rocco, and P Pelosi. Personalized mechanical ventilation settings: Slower is better! In *Annual Update in Intensive Care and Emergency Medicine 2022*, pages 113–127. Springer, 2022.

[Simone *et al.*, 2026] Lorenzo Simone, YF Ferrari Chen, Yves A Lussier, Peter Elkin, and Xinxin Zhu. Ai-driven advances in personalized therapeutic strategies for precision medicine. In *Applied Artificial Intelligence for Drug*

Discovery: From Data-Driven Insights to Therapeutic Innovation, pages 789–816. Springer, 2026.

[Skafte *et al.*, 2019] Nicki Skafte, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[Tang and Wiens, 2021] Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pages 2–35. PMLR, 2021.

[Umerenkov *et al.*, 2025] Dmitry Umerenkov, Alexandr Nesterov, Vladimir Shaposhnikov, Ruslan Abramov, Nikolay Romanenko, Vladimir Kokh, Marina Kirina, Anton Abrosimov, Dmitry V Dylov, and Ivan Oseledets. Ai diagnostic assistant (aida): A predictive model for diagnoses from health records in clinical decision support systems. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 9880–9889, 2025.

[Yin *et al.*, 2022] Changchang Yin, Ruoqi Liu, Jeffrey Caterino, and Ping Zhang. Deconfounding actor-critic network with policy adaptation for dynamic treatment regimes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2316–2326, 2022.

[Yu *et al.*, 2021] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

[Yuan *et al.*, 2023] Yuyu Yuan, Jinsheng Shi, Jincui Yang, Chenlong Li, Yuang Cai, and Baoyu Tang. Conservative q-learning for mechanical ventilation treatment using diagnose transformer-encoder. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2346–2351. IEEE, 2023.

[Yue *et al.*, 2023] Yang Yue, Rui Lu, Bingyi Kang, Shiji Song, and Gao Huang. Understanding, predicting and better resolving q-value divergence in offline-rl. *Advances in Neural Information Processing Systems*, 36:60247–60277, 2023.

[Zhang *et al.*, 2024] Jing Zhang, Linjiajie Fang, Kexin Shi, Wenjia Wang, and Bingyi Jing. Q-distribution guided q-learning for offline reinforcement learning: Uncertainty penalized q-value via consistency model. *Advances in Neural Information Processing Systems*, 37:54421–54462, 2024.

[Zhang *et al.*, 2025] Haodi Zhang, Jiawei Wen, Jiahong Li, Yuanfeng Song, Liang-Jie Zhang, and Lin Ma. Sspnet: Leveraging robust medication recommendation with history and knowledge. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 9465–9473, 2025.