Benchmarking Reinforcement Learning Algorithms for ICU Ventilator Settings: An Interpretable and Probabilistic Patient Environment for Doctor Agents

Ya-Hsi Chang, Po-Chih Kuo*

Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
timothy@gapp.nthu.edu.tw, kuopc@cs.nthu.edu.tw

Abstract

Mechanical ventilation is essential in intensive care units (ICUs), but prolonged use increases patient risk. Reinforcement learning (RL) offers potential for optimizing ventilator management, yet its clinical adoption is limited by the lack of interpretable and realistic simulation environments. We propose an interpretable and probabilistic patient environment simulator based on action-based k-nearest neighbors and empirical transition probabilities, modeling stochastic state transitions grounded in real ICU data. The simulator supports anomaly detection and provides probabilistic next-state distributions to enhance transparency and safety. Within this environment, we benchmark seven offline RL algorithms under clinically guided reward designs, including five distinct reward function configurations to explore the impact of reward shaping on agent behavior. Our results show that RL agents such as Double DQN and NFQ outperform empirical physician policies in meeting extubation guidelines, especially for high-severity patients. This benchmark enables standardized, interpretable evaluation of RL-based decision support tools for critical care.

Code and extended version —
github.com/timothy0203/aaai-26-rl-benchmark-icu-ventilator-patient-environment

Introduction

Mechanical ventilation (MV) is a cornerstone of critical care, providing life-saving support to patients with respiratory failure in ICUs (Liu et al. 2024; Misseri et al. 2024; Wunsch et al. 2013). Despite its benefits, prolonged MV is linked to severe complications, including airway trauma, ventilator-associated pneumonia, and increased morbidity and mortality (Stivi et al. 2024; Bigatello et al. 2007; Hughes et al. 2012; Esteban et al. 2002). These risks underscore the need for timely and effective ventilator management strategies.

Clinical guidelines, such as those from the University of Pennsylvania, provide criteria for extubation readiness, recommending a respiratory rate (RR) $\leq$ 30, heart rate (HR) $\leq$ 130, and oxygen saturation (SpO_2) $\geq 88\%$ on an inspired oxygen fraction (FiO_2) $\leq 50\%$ (Prasad et al. 2017). Standard practice also emphasizes adequate gas exchange, often defined as $SpO_2 \geq 90\%$ on $FiO_2 \leq 40\%$ (Stivi et al. 2024). However, these protocols require adaptation to individual patient physiology, introducing variability in clinician decision-making. The lack of consensus on optimal weaning strategies, compounded by patient heterogeneity, makes MV management a complex, sequential decision-making challenge (Prasad et al. 2017; Chen et al. 2022).

The advent of large-scale ICU datasets, such as MIMIC-IV and eICU (Johnson et al. 2023; Pollard et al. 2018), has enabled the application of artificial intelligence (AI) and reinforcement learning (RL) to enhance MV management. RL, adept at optimizing sequential decisions under uncertainty, aligns naturally with the dynamic nature of ventilator adjustments (Hengst et al. 2023; Peine et al. 2021; Liu et al. 2024). In RL framework (see Figure 1), a doctor agent observes a patient's state (S_t), selects an action (A_t) such as adjusting FiO_2, and receives a reward (R_t) based on clinical extubation guidelines, with the patient transitioning stochastically to a new state (S_{t+1}) (Prasad et al. 2017).

However, the stochasticity of patient responses, where identical interventions can yield diverse outcomes, presents a major obstacle for developing reliable RL policies (Raghu et al. 2017; Misseri et al. 2024). Applying RL safely requires a high-fidelity, interpretable patient simulator to train and evaluate agents before real-world deployment. Yet, existing simulators often suffer from **deterministic transitions**, **limited transparency**, or rely on **complex black-box models** such as LSTMs or Transformer (Chen et al. 2022; Tamboli et al. 2024), which hinder clinical interpretability and trust.

To address this gap, we propose a realistic and interpretable patient environment simulator for ventilated ICU patients. Leveraging an action-based k-nearest neighbors (KNN) approach, our simulator models state transitions using empirical probabilities derived from real ICU data. Unlike black-box models, our action-based KNN method offers transparency by grounding predictions in similar historical cases, enhancing clinical trust and safety through anomaly detection, as interpretable models could use case-based reasoning for complex domains (Rudin 2019). The simulator outputs probabilistic next-state distributions, reflecting the inherent variability in patient responses.

This work advances MV management through the follow-

*Corresponding author.

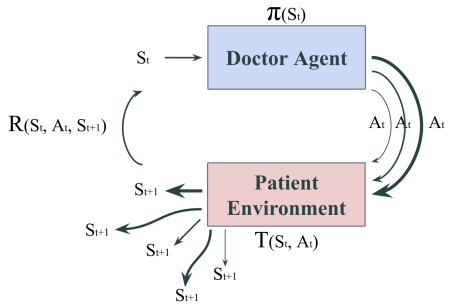

Figure 1: RL framework for ventilator management. A doctor agent observes patient states (e.g., HR, RR, SpO_2), selects an action (e.g., adjusting FiO_2, RR_{set}), and receives a reward based on extubation guidelines. The patient then transitions stochastically to a new state, reflecting physiological variability and physician decision diversity.

ing contributions:

- **Interpretable Patient Environment**: We develop an action-based KNN patient simulator that transparently models stochastic transitions, enabling clinicians to trace simulated outcomes to real-world analogs and detect anomalous RL actions.
- **Doctor Agent RL Benchmarking**: We evaluate seven discrete offline RL algorithms within our environment. The evaluation uses a reward function based on clinical extubation guidelines and patient stability criteria. Performance is benchmarked against an empirical physician policy derived from historical electronic health records (EHR) action distributions.

Related Works

Patient Environment Simulators

Accurate modeling of patient state transitions is foundational for training RL agents in critical care. Existing simulators often leverage complex models such as Long Short-Term Memory networks (LSTMs) (Chen et al. 2022), Recurrent Neural Networks (RNNs) (Raghu et al. 2018), or transformers (Tamboli et al. 2024). For example, (Chen et al. 2022) employed an LSTM to predict vital sign trajectories under MV, while (Raghu et al. 2018) used an RNN to forecast Sequential Organ Failure Assessment (SOFA) scores. These models typically produce deterministic outputs, overlooking the probabilistic nature of patient responses. In contrast, (Li et al. 2022) adopted EHR-derived empirical transition probabilities across 50 states, capturing uncertainty but sacrificing granularity due to coarse state clustering.

Interpretability remains a key challenge. Many simulators, such as those based on LSTMs or transformers, operate as black boxes, obscuring the reasoning behind predictions (Chen et al. 2022; Tamboli et al. 2024). This opacity undermines clinical trust and adoption. Even probabilistic models (Li et al. 2022) struggle with interpretability, as their clustered states lack clear clinical meaning.

RL for MV

RL is increasingly used to optimize MV settings, enabling personalized strategies beyond static protocols. VentAI (Peine et al. 2021) applies Q-learning to adjust PEEP and FiO_2, outperforming standard care in MIMIC-III simulations (Johnson et al. 2016). MHSAC (Chen et al. 2022) addresses hybrid action spaces with a dual-agent approach for efficient control. EZ-Vent (Liu et al. 2024) employs Batch-Constrained Q-learning (BCQ) to mimic clinicians and reduce simulated mortality. Recent models like DeepVent (Kondrup et al. 2022) and X-Vent (Safaei et al. 2024) leverage Conservative Q-Learning (CQL) and explainable AI to improve safety and transparency.

Despite these advances, the field lacks a unified benchmark. Variations in reward design, training pipelines, and evaluation metrics hinder fair comparisons (Schmidt et al. 2023). Many RL systems also rely on simulators with deterministic transitions or poor interpretability, limiting real-world applicability (Peine et al. 2021).

Research Gaps

Current research reveals two major shortcomings:

- **Lack of Interpretability for Patient Environment**: Black-box simulators (e.g., LSTMs, transformers) limit interpretability in decision-making, while clustering approaches oversimplify patient states, diminishing clinical relevance (Chen et al. 2022; Tamboli et al. 2024; Li et al. 2022).
- **No Standardized Benchmark**: The absence of a consistent evaluation framework for RL in MV management hinders progress, as algorithms cannot be reliably compared (Schmidt et al. 2023).

Our work bridges these gaps by introducing an interpretable, probabilistic patient simulator built on real-world data and a standardized RL benchmark for MV optimization. This dual contribution aims to advance both simulation fidelity and algorithmic evaluation in critical care.

Methodology

Figure 2 presents the overall workflow of our study. We first extract the target patient cohort from the MIMIC-IV and eICU databases, followed by preprocessing to obtain the training, test, and out-of-distribution (OOD) evaluation sets. Using the MIMIC-IV training data, we construct the patient environment transition function and derive an empirical physician policy capable of generating simulated trajectories. These components allow us to perform both trajectory-level and transition-level fidelity evaluations. Guided by clinical extubation guidelines, we design several reward function settings and use them to evaluate various discrete offline RL algorithms by simulating their interactions with the patient environment. The entire framework culminates in a ventilator-settings RL benchmark using both the MIMIC-IV test set and the eICU OOD cohort.

Datasets and Cohort Selection

We use two large-scale ICU datasets: MIMIC-IV (Johnson et al. 2023) and eICU-CRD (Pollard et al. 2018). MIMIC-IV

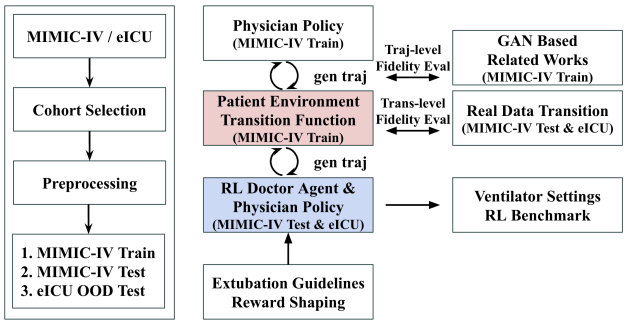

Figure 2: Overview of the proposed methodological framework. The pipeline includes cohort selection, preprocessing, patient-environment construction, physician-policy derivation, fidelity evaluation, and clinically guided reward design, culminating in an RL benchmark.

consists of over 65,000 ICU admissions from a single hospital (2008–2019), while eICU comprises data from over 200 U.S. hospitals (2014–2015). We select adult ICU stays (age $\geq$ 20 years) with at least 24 hours of invasive MV, excluding post-surgical and DNR/DNI patients.

Figure S1 in Supplementary illustrates the cohort selection and variable extraction workflow. The final MIMIC-IV cohort includes 10,633 ICU stays; eICU comprises 3,877.

Dataset Out-of-Distribution (OOD) Analysis

To investigate the impact of an external dataset on patient environment fidelity evaluation and offline agent training, we use the eICU dataset as an out-of-distribution (OOD) source. To quantify distributional shifts between MIMIC-IV and eICU, we conduct a Kolmogorov–Smirnov (KS) test across variables. As shown in Table 1, significant shifts ($p < 0.001$) are observed, particularly in action variables (FiO_2 and RR_{set}), reflecting notable OOD characteristics.

Variable	KS Statistic	p-value
HR	0.035	< 0.001
RR	0.041	< 0.001
SpO_2	0.061	< 0.001
FiO_2	0.100	< 0.001
RR_{set}	0.161	< 0.001

Table 1: KS test comparing variable distributions between MIMIC-IV and eICU datasets. Significant shifts ($p < 0.001$), especially in FiO_2 and RR_{set}, highlight cross-institution variability and OOD characteristics.

Preprocessing Pipeline

We apply a series of preprocessing steps to prepare the data. First, we perform hourly aggregation using severity-aware selection rules for each variable within hourly intervals; full details are provided in the Supplementary Material. Next, we conduct outlier removal to eliminate physiologically implausible values (e.g., HR values outside the range of 30–180 bpm). We then apply missing value imputation using a sample-and-hold approach, where forward-filling is followed by backward-filling. After imputation, any trajectories with missing variables are discarded to ensure data completeness. Finally, we perform discretization of variables based on clinical guidelines: National Early Warning Score (NEWS) thresholds are used for state variables, and empirical bins are applied to FiO_2. This step ensures alignment with domain knowledge and improves interpretability (Liu et al. 2020).

Ventilated Patient Environment: MDP Formulation

We model the patient as a discrete-time Markov Decision Process (MDP). The observation O_t at time t includes the current and two previous hours of history plus baseline:

$$Baseline = (Gender, Age).$$

$$S_t = (HR_t, RR_t, SpO_{2,t}).$$

$$A_t = (FiO_{2,t}, RR_{\text{set},t}).$$

$$O_t = (S_{t-2}, A_{t-2}, S_{t-1}, A_{t-1}, S_t, Baseline)$$

The action $A_t = (FiO_{2,t}, RR_{\text{set},t})$ is applied at time t, and the environment transitions to a new state S_{t+1}.

Action-Based KNN for Transition Modeling

We propose an **Action-Based KNN** to estimate the next-state distribution given (O_t, A_t), as shown in Figure 3:

1. **KNN Retrieval:** Find $K = 100$ nearest neighbors of O_t.
2. **Action Filtering:** Keep neighbors with matching A_t.
3. **Delta Estimation:** Compute average $\Delta S = S_{t+1} - S_t$ from filtered neighbors.
4. **Anomaly Detection:** If no action-matching neighbors are found, label the action as *anomalous*.

This method mimics how clinicians generalize from similar historical cases to predict outcomes.

Alternative Transition Models

We benchmark against the following methods:

- **Keep Current State:** $S_{t+1} = S_t$.
- **Random Walk:** Apply (-1, 0, +1) bin to each variable.
- **Bayesian Smoothing:** Dirichlet-based smoothing of empirical transitions.
- **Gated Recurrent Unit (GRU):** A 2-layer GRU with 128 hidden units and dropout 0.2.
- **Transformer:** A 2-layer Transformer (128 hidden units, 4 heads, 0.2 dropout).

Fidelity Evaluation

We assess environment realism using MMD (Li et al. 2023) at two levels:

- **Transition-level:** MMD between simulated and empirical next-state distributions for each (O_t, A_t) (Figure S2).
- **Trajectory-level:** MMD between generated trajectories and ground-truth sequences from MIMIC-IV.

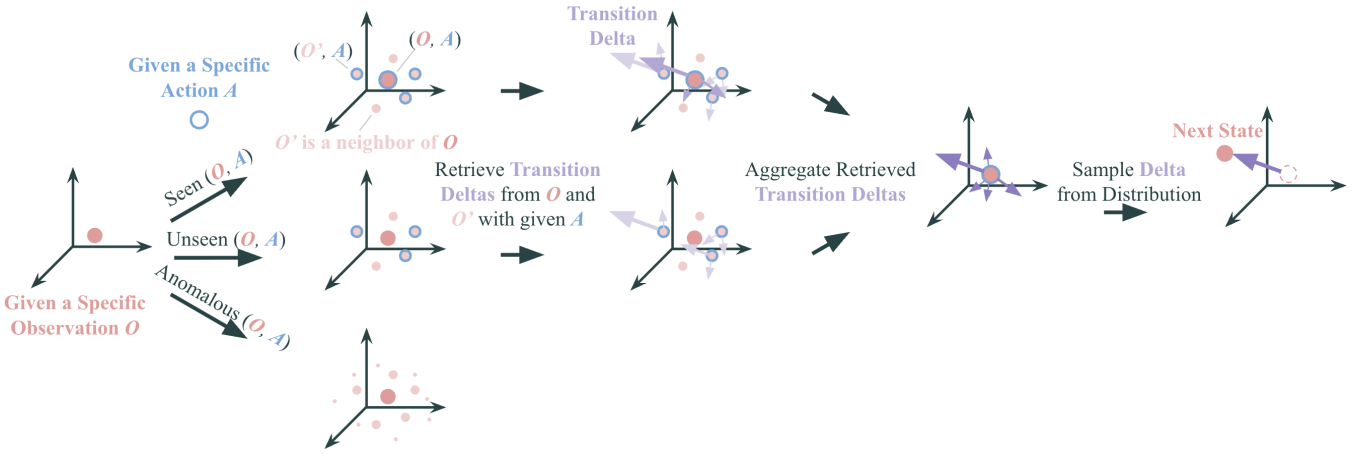

Figure 3: Illustration of Action-Based KNN for Transition Modeling. Pink point represent a specific patient observation O, light pink O' represent neighbors of O, blue frame denote a specific action A, and purple arrows indicate historical transition deltas. Given a current (O, A) pair, the next-state distribution is estimated by aggregating deltas from similar past trajectories. If no neighbors of O exist for the given A, the action A is flagged as anomalous at observation O.

The Gaussian kernel used in MMD is defined as:

$$K(\mathbf{x}, \mathbf{x}') = \sum_i \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma_i^2}\right) \tag{1}$$

where $\mathbf{x}$ and $\mathbf{x}'$ denote sample vectors from two datasets (e.g., simulated vs. real transitions), $\|\cdot\|$ is the Euclidean norm, and σ_i are bandwidth parameters for the Gaussian kernels, selected via the median heuristic over all pairwise distances.

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ represent n samples from the empirical (real) data distribution, and $\mathcal{X}' = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_m\}$ represent m samples from the simulated distribution. The unbiased empirical estimate of the squared MMD is given by:

$$\begin{aligned}\widehat{\mathrm{MMD}}^2(\mathcal{X}, \mathcal{X}') = &\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K(\mathbf{x}_i, \mathbf{x}_j) \\ &- \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} K(\mathbf{x}_i, \mathbf{x}'_j) \\ &+ \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i} K(\mathbf{x}'_i, \mathbf{x}'_j),\end{aligned} \tag{2}$$

Trajectory Comparison with GAN-Based Methods

We generate trajectories by interacting with the patient environment using the physician policy, providing a macro-level evaluation of the environment. These generated trajectories are compared against two baseline trajectory generation methods. The first is **Health Gym** (Kuo et al. 2022), which employs a bidirectional LSTM model trained on MIMIC-III data. The second is **ehrMGAN** (Li et al. 2023), a generative adversarial network that incorporates a VAE-pretrained encoder and a bidirectional LSTM generator.

All models generate 20-step trajectories initialized from the MIMIC-IV training set. We compute MMD to quantify the similarity between generated and real trajectories.

Offline RL Benchmarking

We define a reward function aligned with extubation readiness guidelines. Using this, we train and evaluate 7 discrete offline RL algorithms via `d3rlpy` (Seno and Imai 2021): Behavior Cloning (BC), Neural Fitted Q-Iteration (NFQ), Deep Q-Network (DQN), Double DQN (DDQN), Soft Actor-Critic (SAC), Batch-Constrained Q-Learning (BCQ), Conservative Q-Learning (CQL).

Agents are evaluated based on cumulative rewards and extubation-related metrics in simulated environments, enabling benchmarking of clinical decision policies. Additionally, we assess the impact of five distinct reward shaping designs on agent performance.

Experiments

Patient Environment Configuration

Reward Function Design We define a clinically informed reward function, inspired by (Prasad et al. 2017), to guide agents toward safe and effective ventilator management. The components include:

- **State Reward:** Each vital sign within the clinically recommended range contributes a positive reward of $+\frac{1}{N_{\text{state}}}$, where N_{state} is the number of state variables.
- **State Stability Penalty:** A penalty of $-\frac{1}{N_{\text{state}}}$ is applied if the state changes by two or more bins, discouraging abrupt physiological shifts.
- **Action Stability Penalty:** A penalty of $-\frac{1}{N_{\text{action}}}$ is applied when any action variable changes by more than four bins, discouraging unsafe jumps in ventilator settings.
- **Timestamp Penalty:** A fixed penalty of -1 per timestep encourages timely extubation readiness.
- **Extubation Outcome Reward:** A terminal reward of $+10$ for meeting extubation criteria, and -10 otherwise.

- **Anomalous Action Penalty:** An additional penalty of -10 is given if the selected action is not observed in historical clinical data under similar patient conditions.

Termination Criteria Simulated trajectories are terminated under three conditions: (1) if the patient satisfies extubation guidelines for six consecutive hours (Johns Hopkins All Children's Hospital 2022); (2) if the patient fails to meet the criteria for more than 120 hours; or (3) if the agent takes an anomalous action.

Experimental Workflow

The full experimental pipeline consists of several key stages. First, the Patient Environment is constructed using 70% of the MIMIC-IV as training set, which includes both the transition model and the empirical physician policy. Next, simulator fidelity is evaluated using MMD by comparing against the remaining 30% of MIMIC-IV as test set and the full eICU as OOD test set. Offline RL agents are then trained on the test set and the OOD test set using five different reward design configurations. These trained agents are subsequently evaluated on the Patient Environment, where testing is performed using stratified initial patient states categorized into high, medium, and low severity levels, based on reward distributions from the training trajectories. For each severity category, 100 representative starting states are selected. An empirical physician policy is derived from observed action distributions in training, test, and OOD test set to serve as a baseline. Finally, all policies are benchmarked using six metrics that cover both clinical and behavioral aspects: Total Cumulative Reward, Extubation Meet Rate, Average Trajectory Length, Average Time to Meet Guidelines, Action Diversity, and Anomalous Action Rate.

Reward Design Ablation

To assess how specific reward components influence agent behavior, we define five reward configurations (Table 2). Each configuration selectively activates or deactivates components such as action-stability penalties, state reward / stability penalties, and extubation outcomes.

Design	Act. Pen.	State Rwd.	Extub. Rwd.
Default	w	w	± 10
No Action-Stability Penalty	w/o	w	± 10
No Interm. Reward	w/o	w/o	± 10
High Extub. Reward	w	w	± 100
No Extub. Reward	w	w	0

Table 2: Reward design variants used in offline RL benchmarking. Each configuration activates or removes components such as action stability, intermediate state rewards, and terminal extubation rewards to assess behavioral effects.

Offline RL Agent Setup

We benchmark seven discrete offline RL algorithms using the `d3rlpy` library, all agents use the same hyperparameters to ensure fairness in Supplementary Table S2.

Results

Patient Environment Evaluation

Transition-Level MMD Figure 4 compares several transition models. For seen (O, A) cases (orange and red), the discrete Action-Based KNN achieves the best performance among all methods. For unseen cases (green and purple), GRU slightly outperforms others; however, overall, the Action-Based KNN remains highly competitive compared to GRU and Transformer. Notably, Action-Based KNN offers substantially greater interpretability than these black-box models, making it a practical and transparent choice for clinical decision support. The effect of different K values in Action-Based KNN on MMD is detailed in Figure S3.

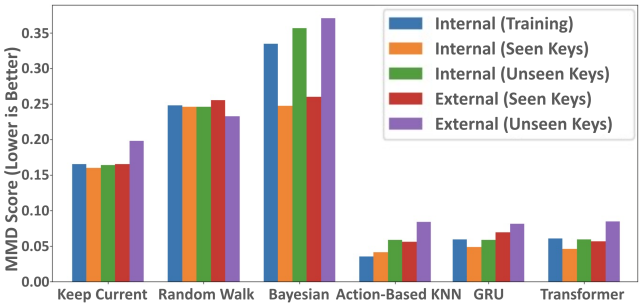

Figure 4: Transition-Level MMD Comparison Across Methods. Each group contains 1,000 trajectories. Detailed values are provided in Table S3.

Trajectory-Level MMD Table 3 presents trajectory-level MMD scores. The physician policy achieves the lowest discrepancy (best match to MIMIC-IV), followed by Health Gym. ehrMGAN exhibits the highest MMD. While Health Gym demonstrates close performance with ours at the trajectory level, it produces implausible transitions—such as fluctuating patient age—highlighting the importance of evaluating both trajectory- and transition-level fidelity.

Method	Trajectory-Level MMD
Physician Policy	**0.228**
Health Gym	0.236
ehrMGAN	0.688

Table 3: Trajectory-level MMD comparison between generated and real MIMIC-IV trajectories. Lower MMD indicates better realism.

Offline RL Benchmarking

Tables S4, S5, and S6 summarize the performance of seven offline RL algorithms compared with the physician policy, across patients of high, medium, and low severity in the training, test (MIMIC), and OOD test (eICU) datasets. Owing to limited space, only the test and OOD test results with high severity are reported in Tables 4 and 5.

Cumulative Reward The physician policy achieves its highest total cumulative reward in the low severity group across all datasets (e.g., 5.25 in training, 3.89 in test, and 4.05 in OOD test).In contrast, rewards are significantly lower for high severity groups (e.g., -16.98 in training). RL-based agents, particularly DDQN, consistently outperform the physician policy in cumulative reward. For example, DDQN achieves 7.1 (train), 7.19 (test), and 8.61 (OOD test medium), highlighting its strong capability in optimizing long-term outcomes.

Extubation Meet Rate The physician policy demonstrates high success in low severity cases (e.g., 92%, 90%, and 87% in training, test, and OOD test, respectively), but its performance degrades notably in high severity groups (50%, 39%, and 37%). Offline RL agents substantially improve extubation meet rates, often achieving $> 92\%$ across all severities. NFQ, DQN, and DDQN reach 100% in medium and low severity groups. However, SAC performs inconsistently—while achieving 94% in high severity (train/test), its performance in OOD test drops drastically (as low as 1–7%), indicating poor generalizability under domain shift.

Time to Meet Extubation Guidelines The physician policy requires longer durations to meet extubation criteria—up to 55.41 hours (high severity, test) and 50.3 hours (OOD test). RL agents demonstrate greater efficiency. DDQN and DQN notably reduce this time across all severity levels, with some reaching less than 17 hours on average. For instance, DDQN achieves 16.02 hours in test-high severity, and DQN achieves 16.7 hours in OOD test-high severity. These improvements suggest better strategic planning and clinical alignment of agent policies.

Action Diversity The physician policies exhibit the highest action diversity (e.g., 58 in train-high severity), due to variability in clinical practice and longer trajectories. Among RL agents, BC retains the most diversity, followed by DDQN and DQN. While RL agents generally show more deterministic decision-making, BC's higher diversity may reflect its supervised learning nature, imitating physician actions more directly.

Anomalous Action Rate The physician policy yields anomalous action rates below 2% across all train/test groups and below 3% in the OOD test set, supporting the suitability of using $K = 100$ in the Action-Based KNN for anomaly detection. RL agents also achieve similarly low rates in most settings, indicating stable learning. However, SAC again shows instability in the OOD test set, with anomalous action rates exceeding 60% in high and medium severity groups—further confirming its lack of robustness under dataset shift.

Summary Overall, offline RL agents outperform the physician policy in cumulative reward, guideline adherence, and efficiency. They consistently achieve higher extubation meet rate and reduce treatment duration while maintaining low anomalous action rates. The physician policy, though diverse and clinically grounded, often requires more time to achieve stabilization in high-risk patients. SAC highlights the need for robust model selection, as its performance varies drastically between datasets. These results establish a strong case for data-driven decision support systems in critical care, provided that reward design and model robustness are carefully addressed.

RL Agent Demonstrations Figure 5 illustrates the first two hours of a high-severity test case. As SpO_2 initially rises, the agent lowers FiO_2, which is followed by a drop in SpO_2. It maintains this level for two hours before reducing FiO_2 further, coinciding with an increased RR and subsequent SpO_2 recovery. As SpO_2 stabilizes within the target range, the agent continues reducing FiO_2. By hour 20, extubation guidelines are met for 6 consecutive hours, indicating a successful ventilator management strategy.

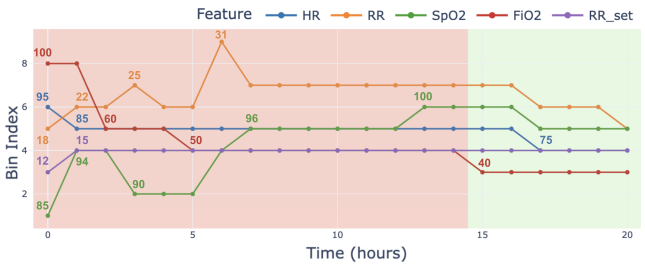

Figure 5: DDQN Agent Trajectory (High Severity, OOD test). The background color reflects guideline compliance—red indicates that extubation guidelines are not met, while green indicates that they are satisfied.

Anomalous Action Example Figure 6 presents a trajectory from an earlier training epoch (BC, test set, 8th epoch). In the final hours, the patient's HR (blue) rises sharply, and SpO_2 (green) drops significantly. While increasing the ventilator RR (purple) seems appropriate, the simultaneous reduction of FiO_2 from 90–100% to 35–40% during low SpO_2 is atypical. As this decision pattern was not observed in historical cases, it is flagged as anomalous.

Figure 6: Anomalous Action Trajectory Example.

Effect of Reward Design Ablation

Tables S7–S27 in Supplementary present the effects of different reward design variants across algorithms. The following discussion outlines the overarching trends observed.

Metric	Phys	BC	NFQ	DQN	DDQN	SAC	BCQ	CQL
Total Reward	-20.07	5.80	7.06	7.09	**7.19**	6.68	6.95	7.16
Meet Rate (%)	39.0	92.0	**95.0**	**95.0**	**95.0**	94.0	94.0	**95.0**
Avg. Len (h)	94.67	19.19	15.84	16.11	**15.42**	15.98	**15.42**	16.54
Time to Meet (h)	55.41	19.72	16.46	16.75	**16.02**	16.74	16.13	17.20
Diversity	**55**	29	11	14	14	15	16	21
Anomalous (%)	**2.0**	8.0	5.0	5.0	5.0	6.0	6.0	5.0

Table 4: High-Severity Results on MIMIC Test Set.

Metric	Phys	BC	NFQ	DQN	DDQN	SAC	BCQ	CQL
Total Reward	-19.38	5.72	6.65	**6.86**	6.77	-19.78	6.46	6.83
Meet Rate (%)	37.0	92.0	**94.0**	**94.0**	93.0	7.0	**94.0**	93.0
Avg. Len (h)	94.29	19.54	18.15	**15.94**	16.26	54.52	18.80	17.03
Time to Meet (h)	50.30	20.12	19.05	16.70	17.18	**13.57**	19.74	18.01
Diversity	**56**	28	20	17	22	37	25	25
Anomalous (%)	**3.0**	8.0	6.0	6.0	7.0	69.0	6.0	7.0

Table 5: High-Severity Results on eICU OOD Test Set.

Action-Stability Penalty Removing the action-stability penalty slightly improved NFQ and SAC performance on test data. For example, NFQ's average reward in the medium severity group increased from 8.49 to 8.64. In the high severity group, NFQ also shortened the time to meet extubation guidelines (from 16.46 to 16.29 hours).

Intermediate vs. Terminal Rewards Removing intermediate rewards encouraged agents to develop clearer long-term goals, reducing trajectory lengths and time to extubation. For example, CQL showed shorter trajectories and faster extubation under no intermediate reward in the high severity test set (15.46 vs. 18.33; 16.19 vs. 19.92) and similarly in OOD test (17.58 vs. 22.32; 18.6 vs. 23.41). In contrast, removing terminal rewards encouraged exploration and increased action diversity, particularly for CQL and BCQ. In OOD test, BCQ (28/30/21) and CQL (26/22/20) achieved the highest diversity across severity groups.

High Terminal Rewards Increasing the terminal extubation reward (10 to 100) reduced trajectory length and time to extubation for DQN, DDQN, and BCQ in test, and DQN, DDQN, and CQL in OOD test. For example, DQN (test-high) improved in length (15.39 vs. 16.76) and time (15.99 vs. 17.15); CQL (OOD test) improved to 16.45 and 17.39. However, this setting increased anomalous actions and then reduced extubation meet rate, indicating a trade-off between aggressive interventions and clinical safety.

Discussion

While the proposed patient environment offers a promising framework for simulating ICU scenarios, it is important to acknowledge its limitations. The primary constraint is the sparsity of training data relative to the vast observation-action space. With an estimated 5×10^{13} possible observations, the 368,169 training pairs cover only a tiny fraction, potentially limiting generalization to unseen states. This sparsity also affects anomaly detection, as rare but valid actions may be misclassified. Additionally, avoiding dimensionality reduction for interpretability restricts scalability to larger feature sets or longer observation windows. Moreover, the limited variable scope—using only a minimal set of clinical features—reduces applicability to real-world settings with more variables.

Nevertheless, unlike black-box simulators such as LSTMs or transformers, our framework is inherently interpretable and allows clinicians to trace decision rationales directly to historical cases. We also deliberately avoid the use of encoders, latent representations, or clustering methods, which, while potentially improving scalability, may obscure clinically meaningful patterns and diminish a physician's intuitive understanding of patient severity. These trade-offs reflect a core design goal: to prioritize transparency and clinical relevance. Future work can build on this foundation by expanding feature coverage, leveraging structured data augmentation, and validating the proposed framework in real-world settings or through collaboration with clinicians.

Conclusion

This work introduces an interpretable, probabilistic patient environment for optimizing MV management using RL. Leveraging an Action-based KNN model trained on MIMIC-IV and eICU data, the framework enables transparent transition modeling by linking actions to similar historical cases and identifying anomalous interventions. Simulator fidelity was validated at both transition and trajectory levels, showing strong alignment with real-world data. Benchmarking across seven offline RL algorithms demonstrated superior performance over empirical physician policies in terms of guideline adherence and time to extubation. The proposed environment offers a clinically relevant and interpretable benchmark for advancing RL in critical care.

References

Bigatello, L. M.; Stelfox, H. T.; Berra, L.; Schmidt, U.; and Gettings, E. M. 2007. Outcome of patients undergoing prolonged mechanical ventilation after critical illness. *Crit Care Med*, 35: 2491–2497.

Chen, S.; et al. 2022. A Model-based Hybrid Soft Actor-critic Deep Reinforcement Learning Algorithm for Optimal Ventilator Settings. *Information Sciences*, 611: 47–64.

Esteban, A.; Anzueto, A.; Frutos, F.; Alía, I.; Brochard, L.; Stewart, T. E.; et al. 2002. Characteristics and outcomes in adult patients receiving mechanical ventilation: a 28-day international study. *JAMA*, 287: 345–355.

Hengst, F. D.; et al. 2023. Guideline-informed Reinforcement Learning for Mechanical Ventilation in Critical Care. *Artificial Intelligence in Medicine*, 147: 102742.

Hughes, C. G.; et al. 2012. Sedation in the intensive care setting. *Clin Pharmacol Adv Appl*, 4: 53.

Johns Hopkins All Children's Hospital. 2022. Invasive ventilation strategies and extubation readiness for premature neonates: Clinical practice guideline. Updated April 26, 2022. Owner: Noura Nickel, MD. Retrieved from Johns Hopkins Medicine website. Accessed 16 Jun. 2025.

Johnson, A. E.; et al. 2016. MIMIC-III, a Freely Accessible Critical Care Database. *Scientific Data*, 3: 1–9.

Johnson, A. E.; et al. 2023. MIMIC-IV, a Freely Accessible Electronic Health Record Dataset. *Scientific Data*, 10: 1–9.

Kondrup, F.; et al. 2022. Towards Safe Mechanical Ventilation Treatment Using Deep Offline Reinforcement Learning. *ArXiv*, 2210.02552.

Kuo, N. I.; et al. 2022. The Health Gym: Synthetic Health-related Datasets for the Development of Reinforcement Learning Algorithms. *Scientific Data*, 9: 1–24.

Li, J.; et al. 2023. Generating Synthetic Mixed-type Longitudinal Electronic Health Records for Artificial Intelligent Applications. *Npj Digital Medicine*, 6: 1–18.

Li, T.; et al. 2022. Electronic Health Records Based Reinforcement Learning for Treatment Optimizing. *Information Systems*, 104: 101878.

Liu, S.; et al. 2024. Reinforcement Learning to Optimize Ventilator Settings for Patients on Invasive Mechanical Ventilation: Retrospective Study. *J Med Internet Res*, 26: e44494.

Liu, V. X.; et al. 2020. Comparison of Early Warning Scoring Systems for Hospitalized Patients With and Without Infection at Risk for In-Hospital Mortality and Transfer to the Intensive Care Unit. *JAMA Netw Open*, 3: e205191.

Misseri, G.; et al. 2024. Artificial Intelligence for Mechanical Ventilation: A Transformative Shift in Critical Care. *Therapeutic Advances in Pulmonary and Critical Care Medicine*, 19: 29768675241298918.

Peine, A.; et al. 2021. Development and Validation of a Reinforcement Learning Algorithm to Dynamically Optimize Mechanical Ventilation in Critical Care. *Npj Digital Medicine*, 4: 1–12.

Pollard, T. J.; et al. 2018. The EICU Collaborative Research Database, a Freely Available Multi-center Database for Critical Care Research. *Scientific Data*, 5: 1–13.

Prasad, N.; et al. 2017. A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units. *ArXiv*, 1704.06300.

Raghu, A.; Komorowski, M.; Celi, L. A.; Szolovits, P.; and Ghassemi, M. 2017. Continuous State-Space Models for Optimal Sepsis Treatment: a Deep Reinforcement Learning Approach. *Proceedings of the 2nd Machine Learning for Healthcare Conference*, 68: 147–163.

Raghu, A.; et al. 2018. Model-Based Reinforcement Learning for Sepsis Treatment. *ArXiv*, 1811.09602.

Rudin, C. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1: 206–215.

Safaei, F.; et al. 2024. X-Vent: ICU Ventilation with Explainable Model-Based Reinforcement Learning. In Endriss, U.; et al., eds., *Proc. of ECAI PAIS 2024*, volume 4719. IOS Press. Open Access, distributed under Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

Schmidt, H.; et al. 2023. Yet Another ICU Benchmark: A Flexible Multi-Center Framework for Clinical ML. *ArXiv*, 2306.05109.

Seno, T.; and Imai, M. 2021. D3rlpy: An Offline Deep Reinforcement Learning Library. *ArXiv*, 2111.03788.

Stivi, T.; Padawer, D.; Dirini, N.; Nachshon, A.; Batzofin, B. M.; and Ledot, S. 2024. Using Artificial Intelligence to Predict Mechanical Ventilation Weaning Success in Patients with Respiratory Failure, Including Those with Acute Respiratory Distress Syndrome. *J Clin Med*, 13: 1505.

Tamboli, D.; Chen, J.; Jotheeswaran, K. P.; Yu, D.; and Aggarwal, V. 2024. Reinforced Sequential Decision-Making for Sepsis Treatment: The PosNegDM Framework With Mortality Classifier and Transformer. *IEEE Journal of Biomedical and Health Informatics*, 28: 3114–3122.

Wunsch, H.; et al. 2013. ICU Occupancy and Mechanical Ventilator Use in the United States. *Critical Care Medicine*, 41: 10.1097/CCM.0b013e318298a139.