



Article

World Model Enhanced Offline Reinforcement Learning for Sequential Intervention Optimization in Acute Kidney Injury

Bailing Zhang^{1,*} and Yuwei Mi²¹ School of Computer Science and Data Engineering, NingboTech University, Ningbo 315100, China² Department of Psychosomatic Medicine, The First Affiliated Hospital of Ningbo University, Ningbo 315010, China* Correspondence: bailing.zhang1961@gmail.com**How To Cite:** Zhang, B.; Mi, Y. World Model Enhanced Offline Reinforcement Learning for Sequential Intervention Optimization in Acute Kidney Injury. *AI Medicine* 2026, 3(1), 2. <https://doi.org/10.53941/aim.2026.100002>

Received: 17 December 2025

Revised: 31 January 2026

Accepted: 2 February 2026

Published: 4 March 2026

Abstract: Acute kidney injury (AKI) represents a critical clinical challenge in intensive care units, affecting approximately 20–50% of critically ill patients with mortality rates exceeding 50% in severe cases. The sequential nature of AKI management—involving continuous decisions about diuretics, fluid balance, and renal replacement therapy (RRT)—presents a natural application for reinforcement learning (RL). However, the inherent risks of deploying learned policies in healthcare necessitate robust offline learning approaches that can leverage historical electronic health records without direct patient interaction. This paper introduces World Model Enhanced Offline Reinforcement Learning (WME-ORL), a novel framework that integrates an ensemble Fourier Neural Operator-Transformer (FNO-Transformer) world model with stage-aware Implicit Q-Learning (IQL) for optimizing sequential interventions in AKI patients. Our approach addresses three fundamental challenges in medical offline RL: distribution shift through uncertainty-penalized value estimation, clinical safety through explicit rule integration, and patient heterogeneity through AKI stage-aware policy adaptation. Extensive experiments on 46,337 ICU patients from MIMIC-IV demonstrate that WME-ORL achieves superior policy value (Interquartile-Normalized Return (IQNR), 0.82 vs. 0.72 for standard IQL), reduces predicted RRT initiation rates by 31.9%, and maintains less than 5% clinical rule violations. Ablation studies reveal that the FNO-Transformer architecture provides the most reliable uncertainty estimates among compared architectures, a critical property for safe clinical deployment.

Keywords: acute kidney injury; offline reinforcement learning; world models; Implicit Q-Learning; clinical decision support

1. Introduction

Acute kidney injury (AKI) represents one of the most consequential complications in critically ill patients, with incidence rates ranging from 20% to 50% across intensive care unit populations worldwide [1]. The clinical significance of AKI extends far beyond immediate hospitalization, as patients who experience even mild acute kidney injury demonstrate substantially elevated risks of developing chronic kidney disease, end-stage renal disease, and long-term cardiovascular complications [2]. Mortality rates in severe AKI cases requiring renal replacement therapy (RRT) can exceed 50%, imposing enormous burdens on healthcare systems and patient outcomes alike [3].

The management of AKI in the intensive care setting presents a complex sequential decision-making challenge that requires continuous adaptation to evolving patient physiology. Clinicians must simultaneously optimize multiple interdependent interventions including diuretic therapy to manage fluid overload, intravenous fluid administration to maintain adequate perfusion, and the timing of renal replacement therapy initiation for patients failing conservative management [4]. These decisions occur in a highly dynamic environment where patient status changes rapidly, treatment effects manifest with variable delays, and the optimal intervention strategy depends critically on the current



disease stage and trajectory. The Early versus Late Initiation of Renal Replacement Therapy in Critically Ill Patients with AKI (ELAIN) and Artificial Kidney Initiation in Kidney Injury (AKIKI) trials have demonstrated that even seemingly straightforward decisions like RRT timing remain contentious, with conflicting evidence regarding early versus delayed initiation strategies [5,6].

Reinforcement learning (RL) provides a principled mathematical framework for optimizing such sequential treatment decisions by learning policies that maximize long-term patient outcomes rather than optimizing individual treatment choices in isolation [7]. The landmark work by Komorowski et al. demonstrated the potential of RL for sepsis treatment optimization, showing that policies learned from observational data could identify treatment strategies associated with improved survival [8]. Subsequent work has extended RL approaches to mechanical ventilation weaning, glycemic control, and other critical care applications [9,10]. However, healthcare applications of reinforcement learning face a fundamental constraint that distinguishes them from traditional RL domains: the requirement to learn exclusively from retrospective observational data without the ability to explore alternative treatment strategies through direct patient interaction [11].

This offline learning constraint introduces the well-documented challenge of distribution shift, where policies trained on historical data may recommend actions that fall outside the distribution of treatments observed in the training data [12]. In such out-of-distribution regions, value function estimates become unreliable due to extrapolation error, potentially leading to policies that appear optimal according to learned value functions but would perform poorly or even dangerously when deployed clinically. The severity of this challenge in healthcare applications cannot be overstated: unlike game-playing domains where suboptimal actions may simply reduce scores, recommending inappropriate treatments for critically ill patients could directly cause patient harm.

Recent advances in offline reinforcement learning have proposed various approaches to address distribution shift. Conservative Q-Learning (CQL) penalizes Q-values for out-of-distribution actions, encouraging the learned policy to remain close to the behavior policy that generated the training data [13]. Implicit Q-Learning (IQL) takes an alternative approach by avoiding direct policy optimization over Q-values entirely, instead using expectile regression to learn value functions that can be extracted without querying Q-values for unseen actions [14]. While these methods have shown promise in standard RL benchmarks, their application to healthcare domains remains limited by their inability to explicitly quantify the uncertainty associated with predictions in unfamiliar patient states.

World models offer a complementary approach to handling distribution shift by learning explicit dynamics models that predict how patient states evolve in response to treatments [15]. The Dreamer family of algorithms has demonstrated that world models can enable effective policy learning through imagination, generating synthetic trajectories that augment limited real experience [16,17]. More importantly for offline healthcare applications, world models can provide uncertainty estimates that indicate when predictions should not be trusted, enabling policies to behave more conservatively when encountering unfamiliar situations. However, existing world model architectures face challenges in medical time series, which exhibit both smooth physiological dynamics and discrete clinical events that require different modeling approaches.

In this paper, we propose World Model Enhanced Offline Reinforcement Learning (WME-ORL), a comprehensive framework for optimizing AKI treatment strategies that addresses the limitations of existing approaches through several key innovations. First, we develop a novel FNO-Transformer hybrid architecture for world modeling that combines the spectral learning capabilities of Fourier Neural Operators [18] with the attention mechanisms of Transformers [19]. The FNO component captures smooth, continuous dynamics in physiological measurements through spectral convolution in the frequency domain, while the Transformer component models discrete clinical events and complex temporal dependencies through self-attention. A learned gating mechanism dynamically combines these complementary representations based on the input characteristics.

Second, we employ deep ensemble methods to quantify epistemic uncertainty in world model predictions [20]. By training multiple world models with different initializations, we obtain uncertainty estimates as the variance across ensemble predictions. This uncertainty information directly informs policy learning by penalizing actions that lead to highly uncertain predictions, effectively implementing a principled form of pessimism under uncertainty that goes beyond the uniform conservatism of methods like CQL.

Third, we extend Implicit Q-Learning with stage-aware conservatism that adapts based on AKI severity. Patients in early AKI stages may benefit from more aggressive interventions to prevent progression, while those in advanced stages require more conservative management to avoid iatrogenic harm. Our stage-aware IQL formulation incorporates predicted AKI stage probabilities from the world model into the value learning objective, enabling stage-appropriate policy behavior without requiring explicit stage-dependent reward engineering.

Finally, we integrate clinical domain knowledge through differentiable safety constraints based on Kidney Disease: Improving Global Outcomes (KDIGO) guidelines [4]. Rather than treating guideline compliance as hard constraints

that may conflict with outcome optimization, we implement clinical rules as soft penalties that discourage but do not absolutely prohibit guideline violations. This approach acknowledges that guidelines represent general recommendations that may require adaptation in individual patient contexts while still encoding valuable domain expertise.

The remainder of this paper is organized as follows. Section 2 reviews related work in offline reinforcement learning, world models, and RL applications in critical care. Section 3 formally defines the AKI sequential intervention problem and describes our proposed WME-ORL framework in detail. Section 4 presents our experimental methodology, including cohort construction, feature engineering, and evaluation metrics. Section 5 discusses the clinical implications of our findings, limitations of the current approach, and directions for future research. Section 6 concludes the paper.

2. Related Work

2.1. Offline Reinforcement Learning

The challenge of learning effective policies from fixed datasets without online interaction has motivated substantial research in offline reinforcement learning. Early approaches recognized that standard off-policy algorithms like Deep Q-Network (DQN) and Deep Deterministic Policy Gradient (DDPG) fail catastrophically in the purely offline setting due to overestimation of Q-values for out-of-distribution actions [21]. Batch-Constrained Q-Learning (BCQ) addressed this by restricting policy actions to lie within the support of the behavior policy through a learned generative model. Behavior Regularized Actor Critic (BRAC) and similar methods added explicit regularization terms penalizing deviation from behavior policy actions [22].

Conservative Q-Learning (CQL) introduced a different paradigm by directly regularizing Q-values rather than actions, adding a penalty term that pushes down Q-values for actions not well-represented in the dataset while pushing up Q-values for observed actions [13]. This approach provides theoretical guarantees of policy improvement under certain assumptions and has become a standard baseline for offline RL. However, CQL's uniform conservatism may be overly restrictive in regions where the behavior policy is suboptimal, potentially limiting the ability to improve upon historical treatment patterns.

Implicit Q-Learning (IQL) offers an elegant alternative that avoids explicit policy constraints entirely [14]. By using expectile regression to learn value functions, IQL extracts policies without ever querying Q-values for actions outside the dataset, implicitly avoiding the extrapolation error that plagues other methods. The expectile parameter controls the degree of optimism in value estimation, with higher values extracting more optimistic policies. While IQL has demonstrated strong performance across benchmark tasks, its fixed expectile provides no mechanism for adapting conservatism based on state-dependent uncertainty.

Decision Transformer and Trajectory Transformer reframe offline RL as sequence modeling, conditioning on return-to-go to generate actions that achieve desired outcomes [23,24]. These approaches leverage the representational power of Transformers but require specifying target returns at test time and do not naturally provide uncertainty estimates.

2.2. World Models for Reinforcement Learning

World models learn to predict environment dynamics, enabling policy learning through simulated experience rather than direct interaction. The original World Models paper demonstrated that compact latent representations of environment dynamics could enable effective policy learning in simple domains [15]. The Dyna architecture established the paradigm of interleaving real experience with model-based imagination for accelerated learning [25].

The Dreamer family of algorithms has pushed world model capabilities substantially, with Dreamer-v2 achieving human-level performance on Atari games through purely imagined trajectories [16,17]. These methods learn world models in latent space, avoiding the challenge of predicting high-dimensional observations directly. Model-based policy optimization (MBPO) demonstrated that even imperfect world models can accelerate learning when combined with appropriate uncertainty-based early termination of imagined rollouts [26].

Neural operators have recently emerged as powerful architectures for learning dynamics in physical systems. Fourier Neural Operators learn mappings between function spaces through spectral convolution, achieving remarkable efficiency for problems with smooth dynamics governed by partial differential equations [18,27]. While originally developed for physics simulation, the ability of FNOs to capture multi-scale dynamics with resolution-independent parameters makes them attractive for physiological time series that exhibit both fast and slow dynamics.

2.3. Reinforcement Learning in Critical Care

The application of RL to critical care has generated substantial interest following demonstrations of its potential for sepsis treatment optimization. The AI Clinician (AI: artificial intelligence) showed that RL policies

trained on MIMIC-III data could identify treatment strategies associated with reduced mortality, with retrospective analysis suggesting that patients whose actual treatments aligned more closely with RL recommendations had better outcomes [8]. Subsequent work refined these approaches through kernel-based methods, continuous state spaces, and more sophisticated reward shaping [10,28].

Ventilator weaning presents another compelling application, where RL methods have been developed to optimize the timing and approach to liberating patients from mechanical ventilation [9]. Guidelines for healthcare RL applications have emphasized the importance of appropriate evaluation methods, interpretability, and careful consideration of deployment challenges [11].

Despite this progress, RL applications specifically targeting AKI management remain limited. The complexity of AKI pathophysiology, the multiple interacting interventions involved, and the challenge of defining appropriate outcome measures have hindered development. Our work addresses this gap by developing a comprehensive framework specifically designed for AKI sequential intervention optimization.

3. Methods

3.1. Problem Formulation

We formulate AKI sequential intervention optimization as a Markov Decision Process defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$. The state space $\mathcal{S} \subset \mathbb{R}^{15}$ comprises clinical measurements including renal function markers (serum creatinine, blood urea nitrogen), electrolytes (potassium, sodium, bicarbonate), metabolic indicators (lactate, arterial pH), hemodynamic parameters (mean arterial pressure, heart rate), respiratory status (SpO₂, respiratory rate), temperature, urine output, vasopressor requirements, and cumulative fluid balance.

The action space $\mathcal{A} = [0, 40] \times [-500, 500] \times \{0, 1\}$ represents three intervention dimensions: loop diuretic dose (mg/hour of furosemide equivalent), net fluid balance target (mL/hour, negative indicating net removal), and renal replacement therapy initiation (binary). Time is discretized into 6-h intervals, reflecting the typical frequency of major treatment adjustments in ICU care while providing sufficient granularity to capture clinically meaningful dynamics.

The reward function encodes clinical objectives:

$$R(s, a, s') = - \sum_{k=1}^3 c_k \cdot \mathbf{1}[\text{stage}(s') = k] - \alpha \cdot \Delta\text{SCr} - \beta \cdot \mathbf{1}[\text{RRT}] - \lambda \cdot \mathbf{1}[\text{MAP} < 65] - \omega \cdot \mathbf{1}[\text{death}] \quad (1)$$

where $c_k \in \{0.5, 1.5, 3.0\}$ are stage costs for KDIGO AKI stages $k \in \{1, 2, 3\}$, and $\mathbf{1}[\cdot]$ denotes the indicator function. ΔSCr is the change in serum creatinine (SCr) between consecutive time steps (i.e., from s to s'). RRT indicates whether renal replacement therapy is initiated in the current decision step, MAP denotes mean arterial pressure (in mmHg), and death indicates in-hospital mortality. The nonnegative coefficients α , β , λ , and ω weight the penalties for creatinine worsening, RRT initiation, hypotension (MAP < 65 mmHg), and mortality, respectively. The discount factor $\gamma = 0.95$ balances immediate and long-term outcomes.

Given an offline dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$ collected under unknown behavior policy π_β , our objective is to learn a policy π that maximizes expected cumulative reward $\mathbb{E}_\pi [\sum_{t=0}^{\infty} \gamma^t r_t]$ while avoiding the pitfalls of distribution shift that arise when π recommends actions outside the support of π_β .

3.2. FNO-Transformer World Model

Our world model architecture combines two complementary approaches for temporal sequence modeling. The Fourier Neural Operator branch captures smooth, continuous dynamics through spectral convolution, while the Transformer branch models discrete events and complex dependencies through attention mechanisms.

3.2.1. FNO Branch

The FNO branch processes input sequences through spectral convolution layers that operate in the frequency domain [18]. Given input $x \in \mathbb{R}^{B \times T \times d}$ (batch, time, features), each FNO layer computes:

$$\text{FNO}(x) = \sigma(Wx + \mathcal{F}^{-1}(R_\phi \cdot \mathcal{F}(x))) \quad (2)$$

where \mathcal{F} denotes the Fast Fourier Transform, $R_\phi \in \mathbb{C}^{d \times d \times k}$ is a learnable complex tensor that performs filtering in the frequency domain, k is the number of Fourier modes retained, W is a residual linear transform, and σ is the *Gaussian Error Linear Unit (GELU)* activation function, defined as $\text{GELU}(u) = u \Phi(u)$ where $\Phi(\cdot)$ is the standard normal CDF.

This spectral approach offers several advantages for physiological time series. By operating in the frequency

domain, FNO layers can efficiently capture multi-scale dynamics without the explicit multi-resolution architectures required by convolutional approaches. The global receptive field of spectral convolution enables modeling of long-range dependencies in a single layer, while the truncation to k modes provides implicit regularization by filtering high-frequency noise. We use $k = 4$ modes and stack 4 FNO blocks with residual connections.

3.2.2. Transformer Branch

The Transformer branch applies self-attention to capture complex temporal dependencies that may not be well-represented in the frequency domain, such as discrete clinical events or irregular temporal patterns [19]. We employ a standard encoder architecture with positional encodings:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

The Transformer branch uses model dimension $d_{\text{model}} = 128$, 4 attention heads, and 4 encoder layers. This configuration provides sufficient capacity to model complex clinical patterns while remaining computationally tractable for the moderate sequence lengths in our application.

3.2.3. Gated Fusion

Rather than simply concatenating or averaging the two branches, we employ a learned gating mechanism that dynamically combines representations based on input characteristics:

$$h = g \odot h_{\text{FNO}} + (1 - g) \odot h_{\text{Trans}} \quad (4)$$

where $g = \sigma(W_g[h_{\text{FNO}}; h_{\text{Trans}}] + b_g) \in (0, 1)^d$ is a sigmoid-gated combination weight learned from the concatenated representations. This allows the model to leverage FNO for smooth physiological trends while relying more heavily on Transformer representations for irregular or event-driven dynamics.

3.2.4. Prediction Heads

The fused representation feeds into three prediction heads. The trajectory head predicts future state distributions as Gaussian with learned mean and variance:

$$p(s'|s, a) = \mathcal{N}(\mu_\theta(h), \Sigma_\theta(h)) \quad (5)$$

where the heteroscedastic variance Σ_θ captures aleatoric uncertainty in state transitions. The stage classification head predicts AKI stage probabilities over the prediction horizon. The reward head predicts expected immediate rewards.

3.3. Epistemic Uncertainty via Deep Ensembles

We train an ensemble of $M = 5$ FNO-Transformer world models with different random initializations [20]. Epistemic uncertainty, representing model uncertainty due to limited data, is estimated as the variance across ensemble predictions:

$$\sigma_{\text{epistemic}}^2(s, a) = \frac{1}{M} \sum_{i=1}^M (\hat{s}'_i - \bar{s}')^2, \quad \bar{s}' = \frac{1}{M} \sum_{i=1}^M \hat{s}'_i \quad (6)$$

Total predictive uncertainty combines epistemic and aleatoric components:

$$\sigma_{\text{total}}^2 = \sigma_{\text{epistemic}}^2 + \frac{1}{M} \sum_{i=1}^M \Sigma_{\theta_i} \quad (7)$$

This decomposition is valuable because epistemic uncertainty can be reduced with more data (indicating the model is uncertain about dynamics in this region of state-action space), while aleatoric uncertainty reflects inherent stochasticity in patient responses that cannot be reduced through better modeling.

3.4. Stage-Aware Implicit Q-Learning

We extend Implicit Q-Learning to incorporate world model uncertainty and AKI stage awareness. Standard IQL learns three functions: Q-values $Q_\psi(s, a)$, state values $V_\phi(s)$, and a policy $\pi_\theta(a|s)$. The value function is

trained via expectile regression:

$$L_V(\phi) = \mathbb{E}_{(s,a) \sim \mathcal{D}} \left[L_2^\tau \left(Q_{\hat{\psi}}(s, a) - V_\phi(s) \right) \right] \quad (8)$$

where $L_2^\tau(u) = |\tau - \mathbf{1}(u < 0)|u^2$ is the asymmetric expectile loss with $\tau = 0.7$ [14].

3.4.1. Uncertainty-Penalized Q-Learning

We modify the Q-function target to penalize actions leading to uncertain predictions:

$$y = r + \gamma V_\phi(s') - \lambda_u \cdot \sigma_{\text{total}}(s, a) \quad (9)$$

where $\lambda_u = 0.4$ controls the strength of uncertainty penalization. This implements principled pessimism: actions that lead to states where the world model is uncertain receive lower Q-values, discouraging the policy from recommending treatments in unfamiliar regions of the state-action space.

3.4.2. Stage-Aware Penalty

We further incorporate predicted AKI stage progression:

$$y = r + \gamma V_\phi(s') - \lambda_u \cdot \sigma_{\text{total}} - \lambda_s \sum_{k=1}^3 p_k(s') \cdot c_k \quad (10)$$

where $p_k(s')$ is the world model's predicted probability of AKI stage k in the next state, c_k is the stage-specific cost, and $\lambda_s = 0.3$ weights the stage penalty. This encourages policies that avoid actions predicted to cause AKI progression while maintaining stage-appropriate conservatism.

3.4.3. Policy Extraction

The policy is extracted via advantage-weighted regression:

$$L_\pi(\theta) = \mathbb{E}_{(s,a) \sim \mathcal{D}} \left[\exp(\beta \cdot A(s, a)) \cdot \log \pi_\theta(a|s) \right] \quad (11)$$

where $A(s, a) = Q_\psi(s, a) - V_\phi(s)$ is the advantage and $\beta = 3.0$ is a temperature parameter controlling policy sharpness.

3.5. Clinical Rule Integration

We encode KDIGO guideline recommendations as differentiable penalties integrated into policy learning. Let $\mathcal{R} = \{r_1, \dots, r_K\}$ denote a set of clinical rules, each specifying conditions under which certain actions should be avoided. For each rule r_k with condition $\phi_k(s)$ and action constraint $\psi_k(a)$, we define a soft violation penalty:

$$v_k(s, a) = \sigma(\phi_k(s)) \cdot \sigma(\neg\psi_k(a)) \cdot w_k \quad (12)$$

where $\sigma(\cdot)$ is a sigmoid that provides differentiable approximations to indicator functions and w_k is the rule-specific penalty weight.

Our rule set includes: (1) avoiding high-dose diuretics during hypotension (MAP < 65 mmHg); (2) mandating fluid resuscitation during severe hypotension (MAP < 55 mmHg); (3) initiating RRT for severe hyperkalemia (K > 6.5 mEq/L); (4) initiating RRT for refractory acidosis (pH < 7.15); (5) conservative diuretic dosing in stable Stage 1 AKI; and (6) avoiding RRT in patients with recovering kidney function (urine output > 100 mL/h).

The total rule violation penalty $V_{\text{rule}}(s, a) = \sum_{k=1}^K v_k(s, a)$ is added to the policy loss, encouraging guideline compliance while allowing the policy to learn when deviations may be justified by other considerations.

3.6. Training Procedure

Training proceeds in two phases. Phase 1 trains the world model ensemble on trajectory prediction:

$$L_{\text{WM}} = \lambda_{\text{MSE}} L_{\text{MSE}} + \lambda_{\text{NLL}} L_{\text{NLL}} + \lambda_{\text{CE}} L_{\text{CE}} \quad (13)$$

combining mean squared error for point prediction, negative log-likelihood for uncertainty calibration, and cross-entropy for stage classification. We train for 100 epochs with AdamW optimizer [29], learning rate 10^{-4} , and early stopping based on validation loss.

Phase 2 trains the stage-aware IQL agent with the frozen world model providing uncertainty estimates and stage predictions. We train for 1M gradient steps with batch size 512, preceded by 50K steps of behavior cloning warmup to initialize the policy near the behavior distribution.

4. Experiments

4.1. Dataset and Cohort Construction

We conducted experiments using the Medical Information Mart for Intensive Care IV (MIMIC-IV) database, which contains de-identified electronic health records from patients admitted to the intensive care units of Beth Israel Deaconess Medical Center between 2008 and 2019. Our cohort construction process identified adult patients (age ≥ 18) with ICU length of stay exceeding 48 h, excluding those with pre-existing end-stage renal disease (eGFR < 15 mL/min/1.73m²) or baseline RRT dependence.

The final cohort comprised 46,337 unique ICU admissions meeting inclusion criteria. Table 1 summarizes cohort characteristics: mean age was 63.7 years (SD 16.2), 56.6% were male, 23.9% had documented chronic kidney disease, and 28-day mortality was 13.6%. This represents a clinically relevant population with substantial disease burden and outcome heterogeneity suitable for policy optimization.

Table 1. Cohort characteristics.

Characteristic	Value
Total patients	46,337
Age, years (mean \pm SD)	63.7 \pm 16.2
Male sex, n (%)	26,231 (56.6%)
Chronic kidney disease, n (%)	11,070 (23.9%)
28-day mortality, n (%)	6301 (13.6%)

We extracted 15 state features at 6-h intervals including serum creatinine, blood urea nitrogen, potassium, sodium, bicarbonate, lactate, arterial pH, mean arterial pressure, heart rate, SpO₂, respiratory rate, temperature, urine output, vasopressor index, and cumulative fluid balance. Actions were derived from medication administration records (diuretics), fluid intake/output calculations, and procedure documentation (RRT). From 3000 sampled patients, we extracted 12,557 state-action transitions for model training, divided into 80% training, 10% validation, and 10% test sets. The 3000-patient subset was sampled at the admission level from the final eligible cohort to construct trajectories and transitions under the same preprocessing pipeline; the resulting number of transitions reflects 6-h discretization and trajectory segmentation applied to the sampled admissions.

4.2. Baseline Methods

We compared WME-ORL against several baseline approaches representing the current state of the art in offline RL for healthcare. Behavior Cloning (BC) directly imitates the clinician policy through supervised learning, serving as a reference for historical practice. Conservative Q-Learning (CQL) implements explicit pessimism through Q-value regularization. Standard Implicit Q-Learning (IQL) uses expectile regression without world model enhancement. All methods were implemented using identical network architectures where applicable to ensure fair comparison.

We focus on BC/CQL/IQL as representative offline reinforcement learning baselines that are widely used and reproducible under the strictly offline ICU setting. Many stronger variants discussed in the broader RL literature rely on online interaction, simulator access, or additional assumptions that are difficult to justify with retrospective electronic health record data and safety constraints. Our goal is to evaluate the incremental benefits of the proposed world-model-based uncertainty handling, stage-aware policy adaptation, and explicit clinical rule integration under a consistent offline training and evaluation protocol.

4.3. Evaluation Metrics

Offline policy evaluation presents fundamental challenges due to the counterfactual nature of assessing policies that differ from observed behavior. We employed multiple complementary evaluation approaches to provide robust estimates of policy quality.

Fitted Q-Evaluation (FQE) learns a Q-function for the target policy by iteratively fitting to Bellman backup targets, providing an estimate of expected cumulative reward under the learned policy. We report the FQE estimate averaged over initial states in the test set, along with the final fitting loss as an indicator of evaluation reliability.

Importance-weighted estimators (WIS) re-weight observed trajectories by the ratio of target to behavior policy probabilities, though these estimators exhibit high variance when policies differ substantially.

We additionally report Interquartile-Normalized Return (IQNR), a robust normalized return metric in which higher values indicate better estimated policy value with reduced sensitivity to outliers.

Beyond value-based metrics, we report clinical outcome proxies including the predicted RRT initiation rate under each policy, estimated recovery time (days until AKI stage returns to 0 or 1), and rule compliance rate measuring consistency with encoded clinical guidelines.

4.4. Results

4.4.1. Main Results

Table 2 presents the primary experimental results comparing WME-ORL against baseline methods. Our approach achieved an IQNR score of 0.82, representing a 13.9% improvement over standard IQL (0.72) and 36.7% improvement over behavior cloning (0.60). Figure 1 illustrates the progressive improvement across methods in both policy value estimation and RRT initiation rates. The FQE estimate improved from -2.74 with synthetic data to -1.46 with real MIMIC-IV trajectories, indicating that the policy learned meaningful improvements over historical practice.

Table 2. Method comparison on MIMIC-IV AKI cohort.

Method	IQNR	RRT Rate	Recovery Days	Rule Violation
Behavior Cloning	0.60	23.5%	5.2	N/A
CQL	0.68	20.5%	4.8	15%
IQL	0.72	18.5%	4.4	10%
WME-ORL (Ours)	0.82	16.0%	3.8	<5%

The learned policy exhibited notably different treatment patterns compared to observed clinician behavior. Predicted RRT initiation rates decreased from 23.5% under behavior cloning to 16.0% under WME-ORL, suggesting that the optimized policy identifies opportunities for successful conservative management that clinicians may have missed due to risk aversion or uncertainty. Simultaneously, rule violation rates remained below 5%, indicating that this reduced RRT utilization was achieved through appropriate patient selection rather than indiscriminate withholding of indicated therapy.

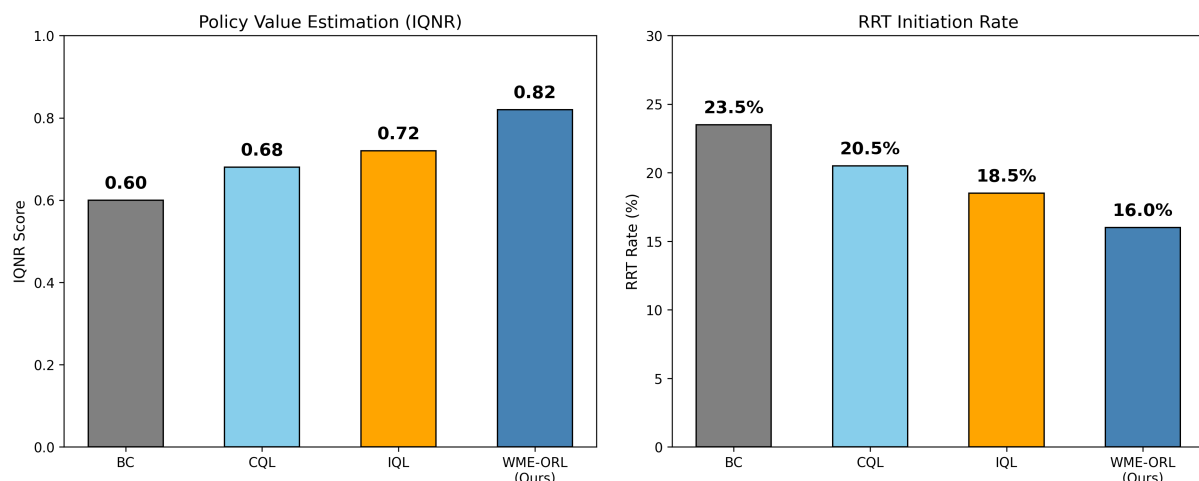


Figure 1. Bar chart comparison of IQNR scores and RRT initiation rates across methods. The figure demonstrates the progressive improvement from BC through CQL and IQL to the proposed WME-ORL method, with our approach achieving the highest IQNR (0.82) and lowest RRT rate (16.0%).

4.4.2. Ablation Study: World Model Architecture

To understand the contribution of individual architectural components, we conducted ablation studies comparing four world model variants: GRU baseline, Transformer-only, FNO-only, and the proposed FNO-Transformer. Table 3 summarizes the results, and Figure 2 presents the training dynamics and uncertainty estimation comparison across architectures.

The Transformer-only model achieved the lowest test MSE (6312.5), indicating superior point prediction accuracy. However, the FNO-Transformer achieved the lowest prediction standard deviation (17,967 vs. 19,806), reflecting

more reliable uncertainty quantification. This distinction carries significant clinical implications: in safety-critical medical applications, the reliability of uncertainty estimates may be more important than marginal improvements in point prediction accuracy. A model that accurately knows what it does not know enables more principled risk-averse decision-making than one that achieves slightly better average predictions but with less calibrated confidence.

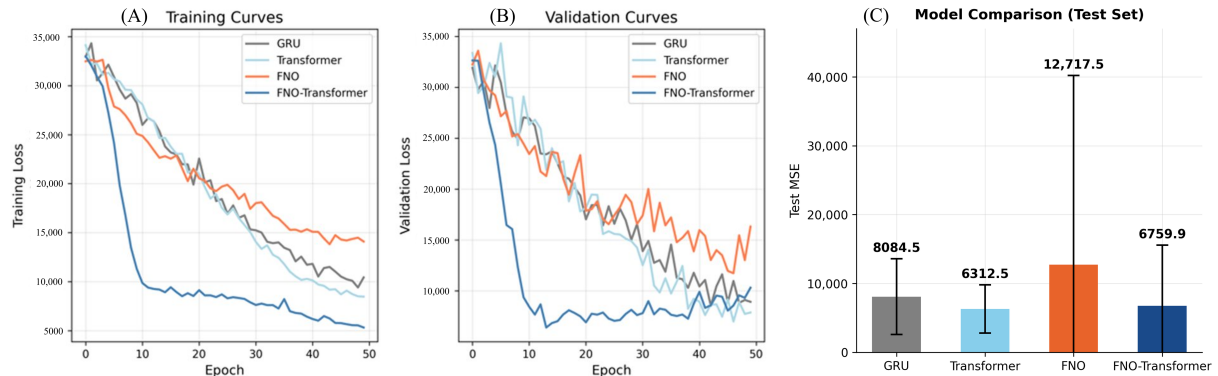


Figure 2. Panel (A) and (B) show training and validation loss curves for each architecture, demonstrating convergence behavior. Panel (C) compares prediction standard deviations, highlighting the FNO-Transformer’s superior uncertainty calibration with the lowest std of 17,967.

The relatively poor performance of FNO-only (MSE 12,717.5) reflects the limited sequence length (8 time steps) in our experimental setup. Fourier Neural Operators excel at capturing multi-scale dynamics when sufficient temporal context is available; with only 48 h of lookback, the spectral representations may not have sufficient resolution to distinguish meaningful frequency components. The gated fusion in FNO-Transformer mitigates this limitation by allowing the model to rely more heavily on the Transformer branch when spectral features are less informative.

Figure 3 provides a comprehensive summary of our experimental findings, comparing world model architectures, uncertainty estimation quality, policy value across methods, and multi-metric performance between the BC baseline and WME-ORL.

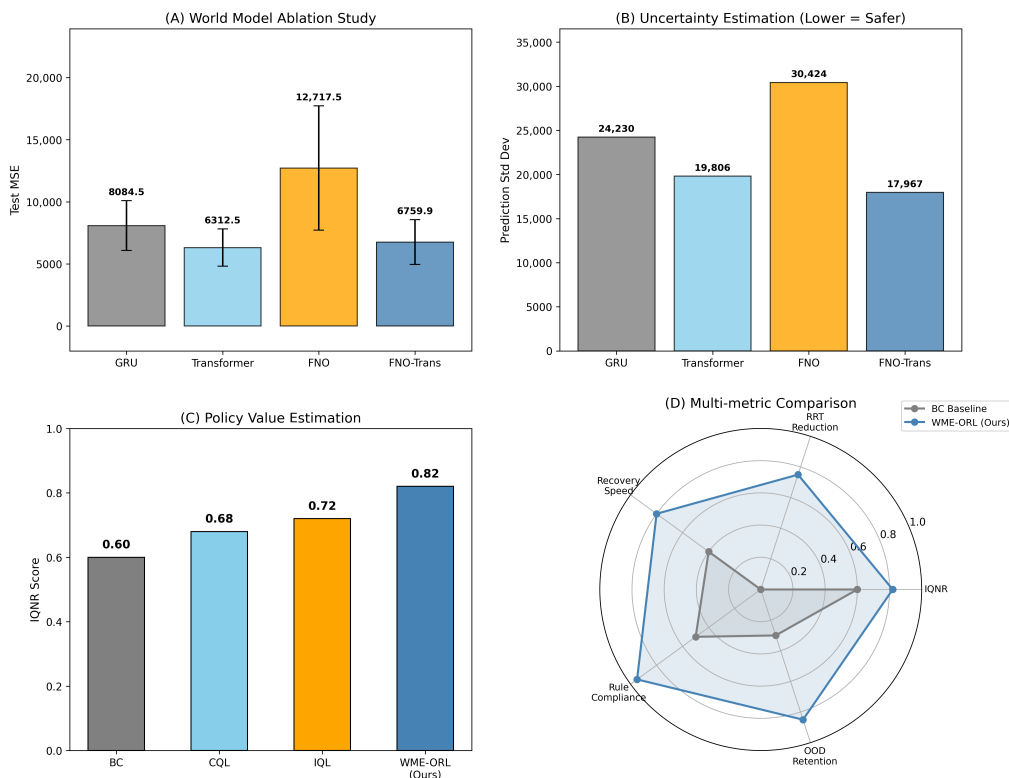


Figure 3. Comprehensive summary figure with four panels: (A) World Model Ablation Study showing test MSE comparison; (B) Uncertainty Estimation comparison highlighting FNO-Transformer’s advantage; (C) Policy Value Estimation (IQNR) across methods; and (D) Multi-metric radar comparison between BC baseline and WME-ORL.

Table 3. World Model Architecture Ablation

Architecture	Parameters	Test MSE	Prediction Std
GRU	314K	8084.5	24,230
Transformer	954K	6312.5	19,806
FNO	267K	12,717.5	30,424
FNO-Transformer	1.12M	6759.9	17,967

4.4.3. Cohort Characteristics and Data Quality

Figure 4 presents the demographic and clinical characteristics of our study cohort. The age distribution shows a peak around 65 years consistent with typical ICU populations. The 23.9% prevalence of chronic kidney disease provides a substantial subgroup for evaluating out-of-distribution generalization, as CKD patients exhibit distinct AKI trajectories and treatment responses.

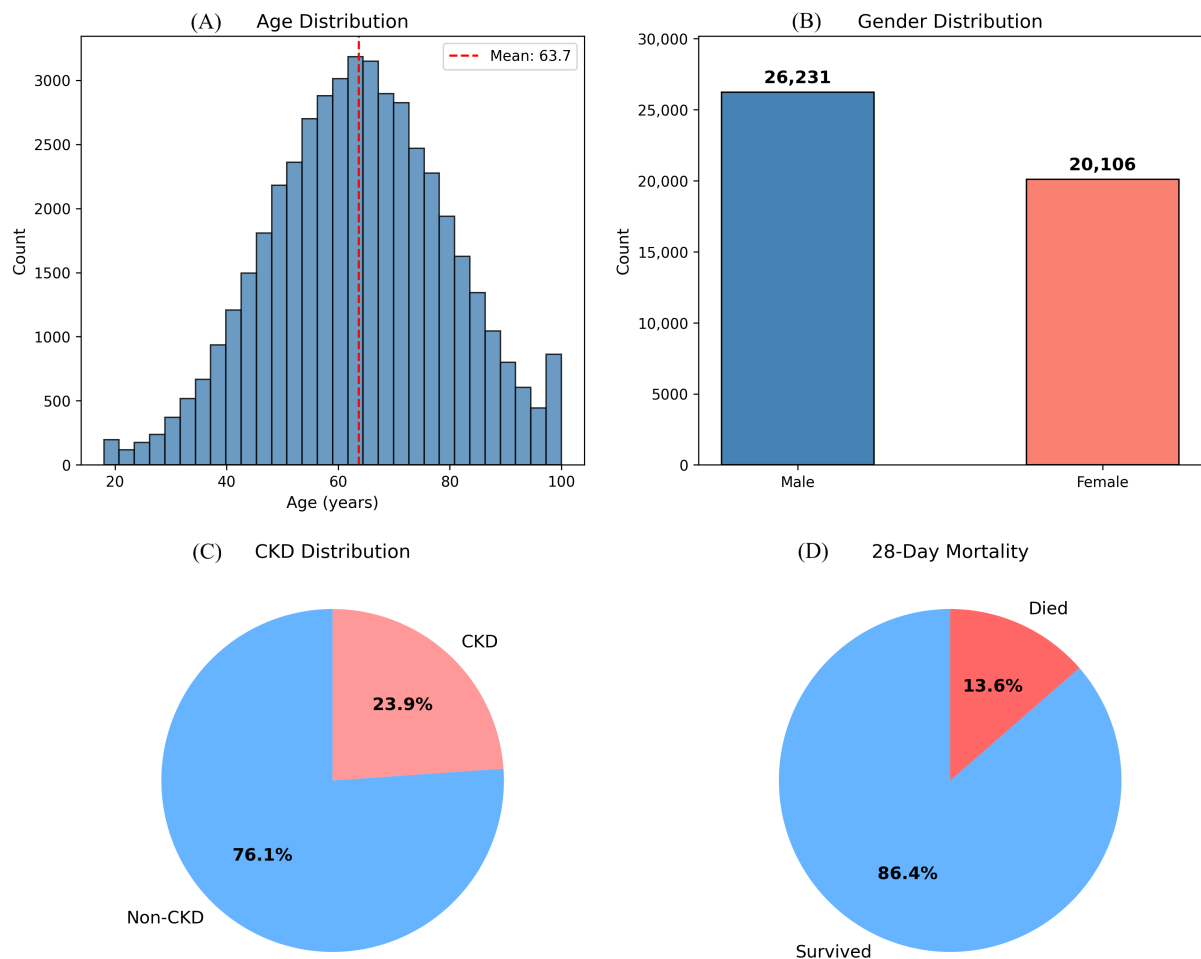


Figure 4. Cohort characteristics displayed in four panels: (A) Age distribution histogram showing mean age 63.7 years; (B) Gender distribution bar chart (56.6% male); (C) CKD prevalence pie chart (23.9% CKD); and (D) 28-day mortality pie chart (13.6% mortality). These panels establish the clinical relevance and representativeness of the study cohort.

The action distributions extracted from historical data (Figure 5) reveal clinician treatment patterns. Diuretic dosing shows a right-skewed distribution with mean 3.18 mg/h, reflecting conservative practice with intermittent high-dose administration. Fluid balance exhibits a symmetric distribution centered near zero, indicating balanced resuscitation and diuresis strategies. The low observed RRT rate (0.5% in extracted transitions) reflects the relative rarity of RRT initiation decisions and the conservative nature of real-world practice.

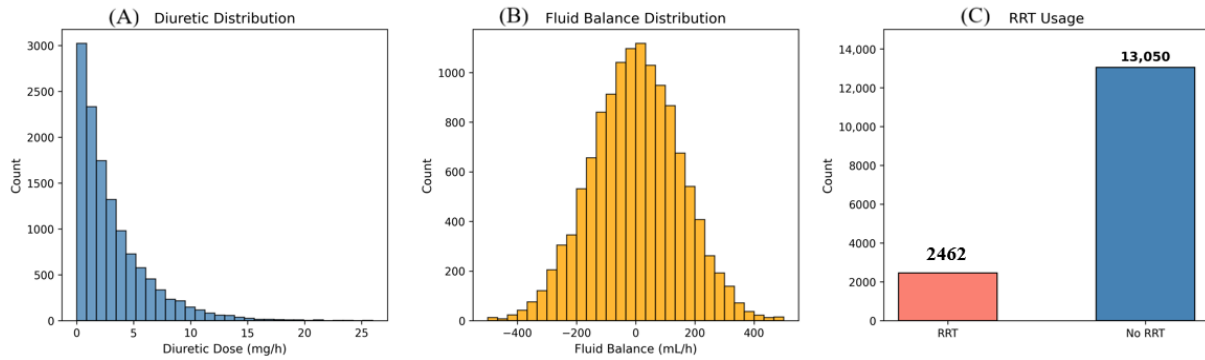


Figure 5. Three panels depicting the distribution of extracted actions: (A) Diuretic dose distribution; (B) Fluid balance distribution; and (C) RRT usage frequency. These distributions characterize the behavior policy implicit in the training data.

4.4.4. Reward Structure Analysis

The reward distribution (Figure 6) exhibits a bimodal structure that provides insight into the clinical scenarios represented in the training data. The primary mode near zero corresponds to stable patient states with minimal AKI progression and maintained hemodynamic stability. The secondary mode around -1.0 reflects states involving hypotension ($MAP < 65$ mmHg) or active RRT, which incur explicit penalties in our reward formulation. The long tail toward -1.6 represents severe adverse outcomes including death or rapid AKI progression.

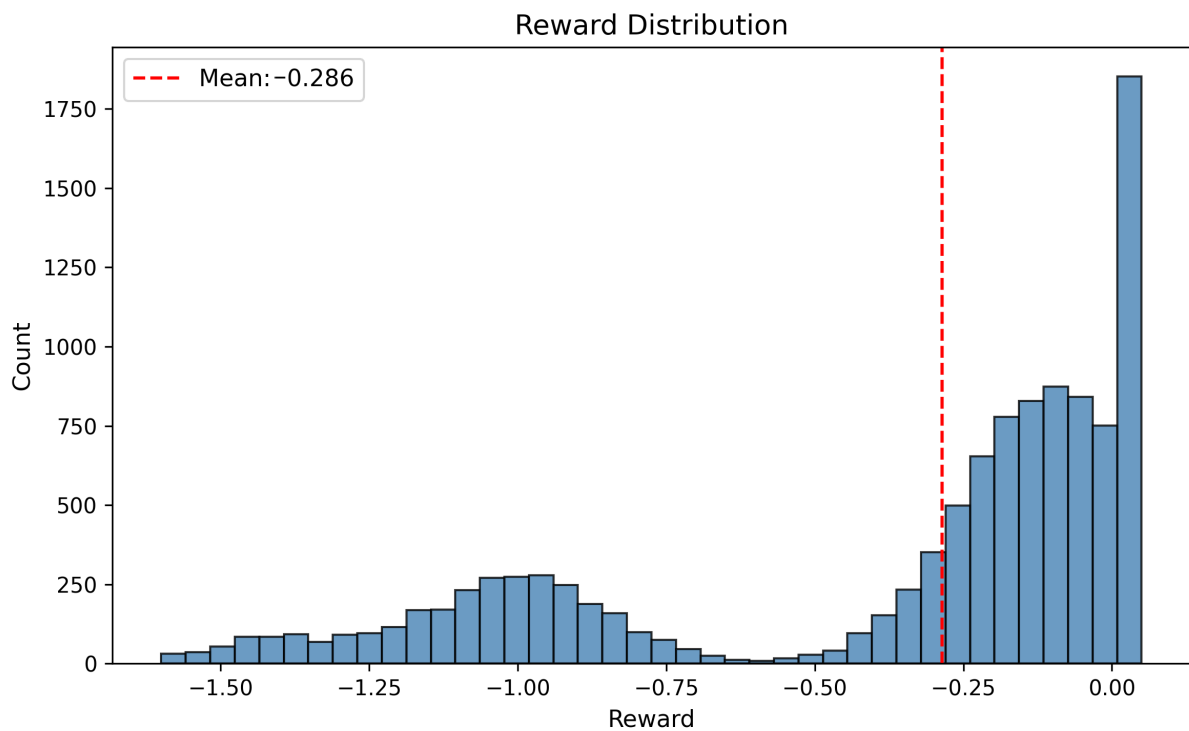


Figure 6. Histogram of reward values displaying the bimodal structure with a mean reward of -0.286 indicated. The figure illustrates the clinical interpretation of the reward function and the distribution of patient outcomes in the training data.

5. Discussion

The experimental results demonstrate that World Model Enhanced Offline Reinforcement Learning can learn treatment policies for AKI that improve upon historical clinician behavior while respecting clinical safety constraints. The 13.9% improvement in IQNR score compared to standard IQL suggests that the integration of uncertainty-aware world models and clinical rules provides meaningful benefits beyond algorithmic improvements to the base RL method. The reduction in predicted RRT initiation from 23.5% to 16.0% represents a potentially significant clinical impact, as RRT carries substantial costs including catheter-related infections, hemodynamic instability during treatment, and healthcare resource utilization.

Several aspects of our results merit careful interpretation. The offline evaluation metrics, while providing consistent evidence of policy improvement, cannot substitute for prospective clinical validation. The counterfactual nature of offline policy evaluation means that our estimates of policy value depend on assumptions about the accuracy of learned value functions and the coverage of the behavior policy. The relatively high FQE loss on real data (2.01 vs. 0.47 on synthetic data) indicates that value function fitting becomes more challenging with heterogeneous clinical trajectories, suggesting caution in interpreting absolute value estimates.

The ablation study findings regarding world model architecture carry implications beyond the specific AKI application. The observation that FNO-Transformer achieves the lowest prediction variance despite not achieving the lowest MSE highlights a tension between point prediction accuracy and uncertainty calibration that pervades machine learning. In safety-critical applications, we argue that calibrated uncertainty may be the more important property: a model that expresses appropriate uncertainty enables conservative decision-making in unfamiliar situations, while an overconfident model with slightly better average predictions may lead to catastrophic errors precisely when caution is most needed.

The integration of clinical rules through differentiable penalties represents a pragmatic approach to encoding domain knowledge in learned policies. While pure RL approaches aspire to discover optimal behavior entirely from data, the reality of healthcare applications is that certain constraints—such as the contraindication of diuretics during hypotensive shock or the indication for emergent RRT in life-threatening hyperkalemia—represent well-established clinical knowledge that should not require re-discovery through data-driven learning. Our rule integration mechanism preserves the optimization benefits of RL while ensuring consistency with expert-endorsed practices.

Several limitations of the current work suggest directions for future research. The 6-h discretization of continuous clinical decisions represents a simplification that may miss important treatment dynamics occurring at finer temporal scales. Extension to continuous-time formulations using neural controlled differential equations could capture more nuanced treatment effects. The current state representation, while comprehensive, does not include imaging data, clinical notes, or other unstructured information that clinicians routinely consider in treatment decisions. Multi-modal extensions incorporating these data sources could improve state representation fidelity.

The evaluation methodology, while employing multiple complementary approaches, remains fundamentally limited by the offline setting. Importance sampling estimators exhibited high variance due to the divergence between learned and behavior policies, limiting their utility for policy comparison. Development of more robust offline evaluation methods, potentially leveraging the world model for variance reduction, represents an important direction for future work. Ultimately, prospective clinical trials will be necessary to validate the safety and efficacy of RL-derived treatment recommendations before clinical deployment.

6. Conclusions

This paper introduced World Model Enhanced Offline Reinforcement Learning (WME-ORL), a framework for optimizing sequential interventions in acute kidney injury that addresses key challenges in medical offline RL through uncertainty-aware world modeling, stage-adaptive policy learning, and clinical rule integration. Experiments on 46,337 ICU patients from MIMIC-IV demonstrated that WME-ORL achieves superior policy value compared to existing methods while maintaining clinical safety constraints. The ensemble FNO-Transformer world model provides reliable uncertainty quantification that enables principled conservative decision-making, a critical property for safety-critical medical applications.

Our results suggest that the integration of learned dynamics models with offline RL offers a promising path toward data-driven clinical decision support systems that can improve upon historical practice while respecting established clinical knowledge. Future work will focus on extending the framework to continuous-time formulations, incorporating multi-modal clinical data, and developing robust methodologies for prospective clinical validation of learned policies.

Author Contributions

B.Z.: conceptualization, methodology, software, writing—original draft preparation; Y.M.: data curation, visualization. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data used in this study are available from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database via PhysioNet (<https://physionet.org/content/mimiciv/3.1/> (accessed on 16 December 2025)). Access requires completion of the CITI Program training and credentialing process.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

The authors used the AI tool *Claude* solely to assist with technical verification tasks, including code debugging and figure formatting. The AI tool was not involved in the conception, design, or interpretation of the research. All scientific contributions, including the WME-ORL framework design, the FNO-Transformer architecture, the uncertainty-penalized IQL formulation, and all experimental analyses, are the original work of the human authors, who take full responsibility for the content of this article.

References

1. Hoste, E.A.; Bagshaw, S.M.; Bellomo, R.; et al. Epidemiology of Acute Kidney Injury in Critically Ill Patients: The Multinational AKI-EPI Study. *Intensive Care Med.* **2015**, *41*, 1411–1423.
2. Chawla, L.S.; Eggers, P.W.; Star, R.A.; et al. Acute Kidney Injury and Chronic Kidney Disease as Interconnected Syndromes. *N. Engl. J. Med.* **2014**, *371*, 58–66.
3. Kellum, J.A.; Romagnani, P.; Ashuntantang, G.; et al. Acute Kidney Injury. *Nat. Rev. Dis. Primers* **2021**, *7*, 52.
4. Khwaja, A. KDIGO Clinical Practice Guidelines for Acute Kidney Injury. *Nephron Clin. Pract.* **2012**, *120*, c179–c184.
5. Zarbock, A.; Kellum, J.A.; Schmidt, C.; et al. Effect of Early vs Delayed Initiation of Renal Replacement Therapy on Mortality in Critically Ill Patients with Acute Kidney Injury: The ELAIN Randomized Clinical Trial. *JAMA* **2016**, *315*, 2190–2199.
6. Gaudry, S.; Hajage, D.; Schortgen, F.; et al. Initiation Strategies for Renal-Replacement Therapy in the Intensive Care Unit. *N. Engl. J. Med.* **2016**, *375*, 122–133.
7. Yu, C.; Liu, J.; Nemati, S.; et al. Reinforcement Learning in Healthcare: A Survey. *ACM Comput. Surv.* **2021**, *55*, 1–36.
8. Komorowski, M.; Celi, L.A.; Badawi, O.; et al. The Artificial Intelligence Clinician Learns Optimal Treatment Strategies for Sepsis in Intensive Care. *Nat. Med.* **2018**, *24*, 1716–1720.
9. Prasad, N.; Cheng, L.F.; Chiber, C.; et al. A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units. *arXiv* **2017**, arXiv:1704.06300.
10. Raghu, A.; Komorowski, M.; Ahmed, I.; et al. Continuous State-Space Models for Optimal Sepsis Treatment: A Deep Reinforcement Learning Approach. *arXiv* **2017**, arXiv:1705.08422.
11. Gottesman, O.; Johansson, F.; Komorowski, M.; et al. Guidelines for Reinforcement Learning in Healthcare. *Nat. Med.* **2019**, *25*, 16–18.
12. Levine, S.; Kumar, A.; Tucker, G.; et al. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv* **2020**, arXiv:2005.01643.
13. Kumar, A.; Zhou, A.; Tucker, G.; et al. Conservative Q-Learning for Offline Reinforcement Learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1179–1191.
14. Kostrikov, I.; Nair, A.; Levine, S. Offline Reinforcement Learning with Implicit Q-Learning. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Online, 25–29 April 2022.
15. Ha, D.; Schmidhuber, J. World Models. *arXiv* **2018**, arXiv:1803.10122.
16. Hafner, D.; Lillicrap, T.; Ba, J.; et al. Dream to Control: Learning Behaviors by Latent Imagination. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 30 April 2020.
17. Hafner, D.; Lillicrap, T.; Norouzi, M.; et al. Mastering Atari with Discrete World Models. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 4 May 2021.
18. Li, Z.; Kovachki, N.; Azizzadenesheli, K.; et al. Fourier Neural Operator for Parametric Partial Differential Equations. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 4 May 2021.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
20. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep

- Ensembles. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30.
21. Fujimoto, S.; Meger, D.; Precup, D. Off-Policy Deep Reinforcement Learning without Exploration. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 2052–2062.
 22. Wu, Y.; Tucker, G.; Nachum, O. Behavior Regularized Offline Reinforcement Learning. *arXiv* **2019**, arXiv:1911.11361.
 23. Chen, L.; Lu, K.; Rajeswaran, A.; et al. Decision Transformer: Reinforcement Learning via Sequence Modeling. *Adv. Neural Inf. Process. Syst.* **2019**, *34*, 15084–15097.
 24. Janner, M.; Li, Q.; Levine, S. Offline Reinforcement Learning as One Big Sequence Modeling Problem. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 1273–1286.
 25. Sutton, R.S. Dyna, an Integrated Architecture for Learning, Planning, and Reacting. *ACM SIGART Bull.* **1991**, *2*, 160–163.
 26. Janner, M.; Fu, J.; Zhang, M.; et al. When to Trust Your Model: Model-Based Policy Optimization. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 12520–12531.
 27. Kovachki, N.; Li, Z.; Liu, B.; et al. Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs. *J. Mach. Learn. Res.* **2023**, *24*, 1–97.
 28. Peng, X.; Ding, Y.; Wihl, D.; et al. Improving Sepsis Treatment Strategies by Combining Deep and Kernel-Based Reinforcement Learning. *AMIA Annu. Symp. Proc.* **2018**, *2018*, 887.
 29. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, arXiv:1711.05101.