Performance and Interpretability Trade-offs in Reinforcement Learning for Sepsis Treatment: Comparing Offline and Online Approaches

Zhiyu Cheng[‡,*], Yalun Ding[1], Chuanhui Peng[1]
[1]Department of Statistics, George Washington University
[*]Corresponding author: zhiyu.cheng@email.gwu.edu

October 28, 2025

Abstract

Sepsis remains a leading cause of mortality in critical care, and reinforcement learning (RL) offers a promising route to data-driven treatment policies. Yet clinical adoption is impeded by the prevailing assumption that interpretability inevitably compromises performance, and that online RL methods necessarily outperform offline approaches. We interrogate these trade-offs by comparing three offline RL methods (Behavior Cloning, Conservative Q-Learning, and Deep Q-Network trained on static datasets) with three online RL methods (Double DQN with Attention, Double DQN with Residual connections, and Soft Actor-Critic with environment interaction) using the gym-sepsis simulator—a MIMIC-III–derived environment for sepsis treatment. Policy quality is assessed through patient-survival rates stratified by Sequential Organ Failure Assessment (SOFA) states, while interpretability is quantified with Linearly Estimated Gradients (LEG), a model-agnostic feature-importance method.

Across 500 evaluation episodes, online RL achieves marginally higher overall survival (95.4% for DDQN-Attention vs. 94.0% for CQL), with a 2.0 percentage point advantage on high-severity patients (90.5% vs. 88.5%). However, comprehensive LEG analysis across all six algorithms reveals a measurable performance-interpretability trade-off: CQL achieves maximum saliency of 40.06 (excellent interpretability with clinically coherent emphasis on blood pressure and lactate), online methods achieve 1.17–3.57 (moderate interpretability), and vanilla DQN achieves only 0.069 (poor interpretability)—spanning three orders of magnitude. Critically, DDQN-Attention's 1.4 percentage point survival advantage over CQL comes at an 11-fold loss in interpretability (3.57 vs. 40.06), demonstrating that the modest performance gain requires sacrificing explainability and incurring extensive environment interaction during training—both infeasible in clinical settings where patient safety and regulatory approval demand transparent decision logic. These results challenge the assumption that interpretability inevitably compromises performance, showing instead that conservative offline RL (CQL) offers the optimal balance for clinical deployment: competitive survival rates with superior explainability and no patient risk during training.

Keywords: Reinforcement Learning, Sepsis Treatment, Interpretability, Conservative Q-Learning, LEG Analysis, Offline RL, MIMIC-III

[*]Equal contribution. Authors listed alphabetically. The authors gratefully acknowledge the support of STAT 8289 - Reinforcement Learning course at George Washington University. This work uses the MIMIC-III database (Johnson et al., 2016) and builds upon the gym-sepsis environment (Raghu et al., 2017).

1 Introduction

Sepsis remains one of the most pressing challenges in critical care, responsible for nearly twenty percent of global mortality and more than $62 billion in annual U.S. healthcare expenditures (Rudd et al. 2020, Fleischmann et al. 2016). Despite successive iterations of the Sepsis-3 definition (Singer et al. 2016) and aggressive early-intervention campaigns, outcomes have plateaued: mortality still ranges from 10–20% for sepsis without shock to 40–50% for septic shock. Clinicians must synthesize heterogeneous physiological signals and act within hours, yet existing protocols offer only population-level heuristics for fluid resuscitation, vasopressor titration, and escalation to organ support (Rhodes et al. 2017). Large randomized trials that re-evaluated early goal-directed therapy (EGDT) (Rivers et al. 2001, ARISE Investigators et al. 2014) underscore the difficulty of prescribing universally optimal intervention thresholds.

Reinforcement learning (RL) has emerged as a candidate framework for tailoring sepsis therapy to patient trajectories. By optimizing long-horizon rewards, RL-based policies can, in principle, balance competing short-term hemodynamic targets against downstream survival. Early work trained Deep Q-Network (DQN) and fitted Q-iteration policies on MIMIC-III data, showing promising retrospective survival estimates (Raghu et al. 2017, Komorowski et al. 2018). However, these studies emphasized expected returns and policy deviations from clinician behavior while offering only qualitative or aggregate descriptions of why certain actions were recommended. As sepsis RL research shifts toward the offline setting—where algorithms must learn exclusively from historical data—questions about policy interpretability become central. Offline methods ranging from Behavior Cloning (BC) to Conservative Q-Learning (CQL) and offline-adapted DQN handle uncertainty and distribution shift differently, which plausibly shapes the transparency of their learned decision rules.

A rigorous understanding of how offline RL algorithms trade off performance and interpretability is still missing. Existing evaluations lack quantitative feature-attribution analyses—a gap with direct regulatory consequences. The U.S. Food and Drug Administration requires explainable AI systems for medical decision support (U.S. Food and Drug Administration 2021), yet no sepsis RL study has demonstrated whether high-performing policies expose clinically plausible decision rationales that clinicians can validate and trust (Holzinger et al. 2017). Moreover, interpretability techniques for sequential decision-making—such as Linearly Estimated Gradients (LEG) saliency maps (Greydanus et al. 2018)—have rarely been applied to healthcare RL, so it remains unclear which algorithmic choices yield explanations aligned with accepted sepsis physiology.

Responding to this gap, we ask: *Can offline RL algorithms for sepsis simultaneously deliver state-of-the-art survival performance and clinically interpretable decision rationales?* We hypothesize that performance–interpretability trade-offs depend on algorithmic design and that conservatism in the objective (as in CQL) can enhance both safety and transparency. To test this hypothesis, we train BC, CQL, and DQN policies on a dataset of simulated patient trajectories generated by rolling out a heuristic policy in the gym-sepsis simulator—an environment whose dynamics and outcome models were trained on MIMIC-III data (Raghu et al. 2017). We jointly evaluate survival outcomes stratified by Sequential Organ Failure Assessment (SOFA) scores and LEG-based feature saliency, ensuring that both performance and interpretability are quantified rigorously.

This study contributes (i) the first quantitative benchmark of offline RL algorithms on both outcome metrics and interpretability for sepsis management, (ii) empirical evidence that the presumed performance–interpretability tension is not inevitable, with CQL matching the

survival of alternative policies while providing salient, guideline-consistent explanations, and (iii) methodological guidance on applying LEG analysis to safety-critical RL deployments. Section 2 surveys clinical RL and interpretability research, Section 3 formalizes the sepsis decision process and LEG framework, Section 4 details experimental design, Section 5 reports performance and interpretability findings, Section 6 interprets the implications for clinical adoption, and Section 7 outlines future research directions.

2 Related Work

Our work builds on three research areas: reinforcement learning for sepsis treatment, offline RL algorithms, and interpretability methods for RL policies.

2.1 Reinforcement Learning for Sepsis Treatment

Raghu et al. (2017) pioneered deep RL for sepsis treatment using the MIMIC-III database, formulating treatment as a discrete-action MDP with a 5×5 action grid (IV fluid $\times$ vasopressor dosing). Their work established the gym-sepsis simulation environment we use in this study. Komorowski et al. (2018) developed the AI Clinician using fitted Q-iteration, achieving 98% survival in retrospective simulation. While these studies demonstrated high performance, **neither provided quantitative feature-attribution analysis**. Their interpretability assessments relied on visualizing aggregate action distributions, revealing *what* the policy does but not *why*—a critical gap for regulatory approval and clinical trust. Recent work by Yao et al. (2021) applied CQL to sepsis but did not evaluate policy interpretability systematically.

2.2 Offline Reinforcement Learning

Offline RL learns from fixed datasets without environment interaction, addressing distributional shift when learned policies select out-of-distribution (OOD) actions (Levine et al. 2020). **Behavior Cloning (BC)** (Pomerleau 1991) treats offline RL as supervised imitation learning, avoiding distributional shift but cannot improve beyond the behavioral policy. **Conservative Q-Learning (CQL)** (Kumar et al. 2020) adds a conservatism penalty that discourages high Q-values for OOD actions, providing safety guarantees for healthcare. CQL's penalty encourages simpler Q-function representations aligned with the behavioral policy—when the behavioral policy follows interpretable threshold rules, CQL may learn Q-functions with strong gradients detectable by saliency analysis. **Deep Q-Network (DQN)** (Mnih et al. 2015) combines Q-learning with deep networks and experience replay; originally designed for online learning, it can be adapted to offline settings but tends to overestimate OOD action values.

2.3 Interpretability Methods in RL

Regulatory agencies require explainable AI for medical decision support (Holzinger et al. 2017), yet deep RL policies are notoriously opaque. Greydanus et al. (2018) introduced **Linearly Estimated Gradients (LEG)**, a perturbation-based method that approximates Q-function gradients via local linear regression, producing saliency maps highlighting which features drive action selection. We adopt LEG because it is model-agnostic (enabling fair comparison across algorithms), produces quantitative saliency scores, and aligns with clinical

intuition where clinicians weight physiological indicators. No prior healthcare RL work has systematically compared interpretability across algorithms using gradient-based methods.

2.4 Research Gap

Despite a decade of sepsis RL research achieving strong retrospective performance, quantitative interpretability evaluation remains absent. While offline RL methods differ fundamentally in handling distributional shift, no study has examined whether these algorithmic differences translate to interpretability differences. Our work addresses this gap by jointly benchmarking offline and online RL on both survival outcomes and LEG-based interpretability, providing the first quantitative evidence on the performance–interpretability trade-off across RL paradigms.

3 Problem Formulation

We formulate sepsis treatment as a finite-horizon Markov Decision Process (MDP) in the offline reinforcement learning setting, where the goal is to learn an optimal policy from a fixed dataset without further environment interaction. We define interpretability through the Linearly Estimated Gradients (LEG) framework for quantitative feature importance measurement.

3.1 MDP Formulation

The sepsis treatment MDP is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. The **state space** $\mathcal{S} \subset \mathbb{R}^{46}$ captures patient physiological condition through laboratory values, vital signs, and clinical severity scores as detailed in Section 4.1. The **action space** $\mathcal{A}$ contains 25 discrete actions representing a 5×5 grid of IV fluid and vasopressor dosing levels. The **transition dynamics** $\mathcal{P}(s_{t+1}|s_t, a_t)$ are learned from MIMIC-III data via the gym-sepsis simulator (Raghu et al. 2017). The **reward function** uses sparse terminal rewards: $\mathcal{R}(s_T, a_T) = +15$ for survival, -15 for death, and 0 for intermediate steps. We use discount factor $\gamma = 0.99$.

A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps states to action distributions. The goal is to find $\pi^* = \arg\max_\pi V^\pi(s)$ where the value function is:

$$V^\pi(s) = \mathbb{E}_{\tau \sim \pi, \mathcal{P}} \left[\sum_{t=0}^{H-1} \gamma^t \mathcal{R}(s_t, a_t) \,\middle|\, s_0 = s \right]. \tag{1}$$

The action-value function $Q^\pi(s, a)$ represents expected return when taking action a in state s and following π thereafter. The optimal policy is derived via $\pi^*(s) = \arg\max_a Q^*(s, a)$.

3.2 Offline RL Setting

In offline RL, the agent learns exclusively from a fixed dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$ collected under a behavioral policy, without environment interaction during training (Levine et al. 2020). The central challenge is *distributional shift*: the learned policy π may select out-of-distribution (OOD) actions where Q-value estimates are unreliable due to extrapolation error (Fujimoto et al. 2019).

Conservative Q-Learning (CQL) (Kumar et al. 2020) addresses this via pessimism, penalizing Q-values for OOD actions:

$$\min_Q \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q(s,a)-\left(r+\gamma\max_{a'}Q(s',a')\right)\right)^2\right]+\alpha\cdot\mathbb{E}_{s\sim\mathcal{D}}\left[\log\sum_a\exp(Q(s,a))-\mathbb{E}_{a\sim\pi_\beta}[Q(s,a)]\right], \tag{2}$$

where $\alpha > 0$ controls conservatism strength. Behavior Cloning (BC) avoids distributional shift by imitating the behavioral policy via supervised learning: $\pi_{\text{BC}} = \arg\max_\pi \mathbb{E}_{(s,a)\sim\mathcal{D}}[\log\pi(a|s)]$, but cannot improve beyond the behavioral policy's performance.

3.3 Interpretability via LEG

Interpretability quantifies the extent to which clinicians can understand policy decisions—critical for regulatory approval (U.S. Food and Drug Administration 2021) and clinical trust (Holzinger et al. 2017). We use Linearly Estimated Gradients (LEG) (Greydanus et al. 2018), a model-agnostic perturbation method measuring feature importance.

For a policy π and state s, LEG approximates the saliency (gradient) of the Q-function with respect to each state feature via:

1. Sample $M = 1000$ perturbations $\delta^{(m)} \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.05$
2. Evaluate perturbed Q-values: $\Delta Q^{(m)} = Q(s+\delta^{(m)}, a) - Q(s,a)$
3. Fit linear regression: $\Delta Q^{(m)} \approx \sum_{j=1}^{46} w_j \cdot \delta_j^{(m)}$ via OLS
4. Extract saliency scores: $\text{Saliency}_j(s,a) = w_j$

We quantify interpretability via: (1) **maximum saliency magnitude** $\max_{s,j}|\text{Saliency}_j(s,a)|$, measuring signal strength; (2) **saliency range**, capturing feature differentiation; (3) **clinical coherence**, assessing whether top features align with medical knowledge (e.g., blood pressure, lactate for sepsis). Policies with strong saliency signals (> 10), large ranges, and high clinical coherence are deemed interpretable.

4 Methods

We describe the experimental setup: the gym-sepsis simulation environment, offline and online RL algorithms, LEG interpretability analysis, and evaluation protocol.

4.1 Environment and Data

4.1.1 Gym-Sepsis Simulator

We use Gym-Sepsis (Raghu et al. 2017), an RL simulator for ICU sepsis treatment trained on MIMIC-III data (Johnson et al. 2016).

State Space. At each timestep, the state is a 46-dimensional vector spanning laboratory values (lactate, creatinine, platelet count, etc.), vital signs (blood pressure, heart rate, SpO_2, etc.), demographics (age, gender, race), clinical severity scores (SOFA, LODS, SIRS, qSOFA, Elixhauser), and treatment status (mechanical ventilation, blood culture). The SOFA score (Vincent et al. 1996) ranges from 0–24, with higher values indicating greater organ dysfunction; we use SOFA for severity stratification in Section 4.4.

Action Space. A discrete 5×5 grid over IV fluid and vasopressor dosage bins (action $= 5 \times$ IV_bin + VP_bin), yielding 25 actions.

Episode Dynamics & Reward. Episodes span ICU stays (4-hour timesteps) until discharge or death. Sparse reward: $r_t = +15$ (survival), -15 (death), 0 (intermediate).

4.1.2 Offline Training Dataset

We generated an offline dataset of 10,000 episodes (100K transitions) using a heuristic policy based on clinical guidelines (Rhodes et al. 2017, Seymour et al. 2016), achieving 94.6% survival. Data partitioning: 9,000 train, 500 validation, 500 test episodes.

4.2 Algorithms

We compare three offline RL algorithms representing different learning paradigms: Behavior Cloning (supervised learning), Conservative Q-Learning (offline Q-learning), and Deep Q-Network (online RL adapted for offline evaluation). All algorithms use the same neural network architecture for fair comparison: a 3-layer multilayer perceptron (MLP) with hidden dimensions [256, 256, 128] and ReLU activations.

4.2.1 Behavior Cloning (BC)

Behavior Cloning treats offline RL as a supervised learning problem, training a policy to imitate the behavioral policy by minimizing the negative log-likelihood of observed actions (Pomerleau 1991). Formally, given a dataset $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^{N}$ of state-action pairs, BC learns a policy $\pi_\theta(a|s)$ by solving:

$$\theta^* = \arg\min_\theta -\frac{1}{N}\sum_{i=1}^{N} \log \pi_\theta(a_i|s_i) \tag{3}$$

BC is computationally efficient and stable, but suffers from distribution shift when the learned policy encounters states not well-represented in the offline dataset (Ross & Bagnell 2010).

Implementation. We use d3rlpy's DiscreteBCConfig with batch size 1,024, learning rate 1×10^{-3} (Adam), training for 50,000 gradient steps (10 epochs $\times$ 5,000 steps/epoch).

4.2.2 Conservative Q-Learning (CQL)

Conservative Q-Learning (Kumar et al. 2020) is an offline RL algorithm that learns a conservative Q-function to avoid overestimation on out-of-distribution actions. CQL augments the standard Bellman error with a conservatism penalty that pushes down Q-values for unseen actions while pushing up Q-values for actions in the dataset:

$$\min_Q \alpha \cdot \mathbb{E}_{s\sim\mathcal{D}}\left[\log \sum_a \exp Q(s,a) - \mathbb{E}_{a\sim\pi_\beta} Q(s,a)\right] + \frac{1}{2}\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[(Q(s,a) - \mathcal{T}^\pi Q(s,a))^2\right] \tag{4}$$

where α controls the strength of the conservatism penalty, π_β is the behavioral policy, and $\mathcal{T}^\pi$ is the Bellman operator.

The conservatism penalty encourages the learned Q-function to assign lower values to actions that were not taken by the behavioral policy, reducing the risk of selecting suboptimal actions due to Q-value overestimation. The policy is derived as $\pi(s) = \arg\max_a Q(s,a)$.

Implementation. We use d3rlpy's DiscreteCQLConfig with batch size 1,024, learning rate 3×10^{-4} (Adam), $\alpha = 1.0$, target network updates every 2,000 steps, training for 200,000 gradient steps.

4.2.3 Deep Q-Network (DQN)

Deep Q-Network (Mnih et al. 2015) is a foundational deep RL algorithm that combines Q-learning with deep neural networks. DQN uses two key techniques for stability: (1) experience replay, which stores transitions in a replay buffer and samples mini-batches for training, and (2) a target network Q_{θ^-} that is periodically synchronized with the main network Q_θ to stabilize Q-value targets.

The Q-function is updated to minimize the temporal difference (TD) error:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - \left(r + \gamma \max_{a'} Q_{\theta^-}(s',a')\right)\right)^2\right] \tag{5}$$

The policy is derived greedily as $\pi(s) = \arg\max_a Q_\theta(s,a)$, with ϵ-greedy exploration during training (ϵ annealed from 1.0 to 0.05).

Implementation and Training Paradigm. We use the Stable-Baselines3 library for DQN training. Unlike BC and CQL, which are designed explicitly for offline learning from a fixed dataset, DQN was trained *online* by interacting with the Gym-Sepsis simulator and accumulating experience in a replay buffer of size 100,000. This methodological choice reflects DQN's original design as an online RL algorithm (Mnih et al. 2015) and enables us to compare interpretability across both offline-specific methods (BC, CQL) and online methods adapted for safety-critical domains. We label our study as focusing on "offline RL" because BC and CQL are trained offline, and *all three algorithms are evaluated identically in offline mode*—i.e., policies are tested on a held-out set of 500 episodes without further environment interaction. This evaluation protocol ensures fair comparison: DQN's online training provides it with potentially richer exploration data, yet it must still generalize to unseen test episodes in the same manner as offline-trained policies. Thus, our interpretability analysis reflects how each algorithm's learned representations (whether from offline or online training) manifest in deployment settings where no further learning occurs.

DQN uses batch size 256, learning rate 1×10^{-4} (Adam), target network updates every 1,000 steps, ϵ-greedy exploration ($1.0 \to 0.05$), training for 100,000 timesteps.

4.2.4 Online RL Algorithms

To provide a comprehensive comparison between offline and online RL paradigms, we also evaluate three state-of-the-art online RL algorithms with architectural innovations (implemented by collaborator Y. Ding). Unlike the offline methods above, these algorithms train by interacting with the Gym-Sepsis simulator, collecting 1 million timesteps of experience through exploration. This comparison illuminates the performance-safety trade-off: online methods can explore beyond the behavioral policy's distribution but require environment access during training—a significant constraint in clinical settings where patient safety prohibits trial-and-error learning.

4.2.4.1 Double DQN with Attention (DDQN-Attention). This algorithm extends Double DQN (Van Hasselt et al. 2016) with a multi-head self-attention mechanism in the encoder network. Double DQN addresses Q-value overestimation by decoupling action selection and evaluation: the main network selects the best action, while the target network

evaluates it. The attention layer allows the model to dynamically weight different state features based on their relevance to the current decision:

$$h_t = \text{MultiHeadAttention}(s_t, s_t, s_t) + s_t \tag{6}$$

where the residual connection helps gradient flow during backpropagation. The attention mechanism computes scaled dot-product attention across 4 parallel heads, each learning different feature correlations. The encoder uses two hidden layers of 256 and 128 units respectively, with the attention layer inserted after the first hidden layer to capture high-level feature interactions.

4.2.4.2 Double DQN with Residual Connections (DDQN-Residual). This variant incorporates deep residual networks (He et al. 2016) to enable training of deeper Q-networks without gradient vanishing. The architecture uses three hidden layers of 256 units each, with skip connections between layers:

$$h_{l+1} = \sigma(\text{LayerNorm}(W_l h_l + b_l + h_l)) \tag{7}$$

where σ is the ReLU activation, W_l and b_l are learnable weights and biases, and the additive skip connection h_l preserves gradient information. Layer normalization stabilizes training by normalizing activations within each layer. The residual architecture is hypothesized to learn more complex value functions by decomposing Q-value estimation into a base value plus incremental adjustments.

4.2.4.3 Soft Actor-Critic (SAC). SAC (Haarnoja, Zhou, Abbeel & Levine 2018) is a maximum entropy RL algorithm that optimizes both expected return and policy entropy, encouraging exploration and robustness. The objective function is:

$$J(\pi) = \mathbb{E}_{\tau \sim \pi} \left[\sum_t r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot|s_t)) \right] \tag{8}$$

where $\mathcal{H}(\pi(\cdot|s))$ is the entropy of the policy at state s, and α is a temperature parameter that balances exploitation (maximizing reward) and exploration (maximizing entropy). We use the discrete action space variant of SAC with a residual encoder architecture (3 layers of 256 units with skip connections). The temperature α is automatically tuned during training using a dual gradient descent approach (Haarnoja, Zhou, Hartikainen, Tucker, Ha, Tan, Kumar, Zhu, Gupta, Abbeel & Levine 2018), starting from $\alpha = 0.2$ and adjusting to maintain a target entropy equal to 95% of the maximum entropy $\log(25)$ for the 25-action space.

4.2.4.4 Training Details. All three online RL algorithms were trained with 1,000,000 environment interaction steps using experience replay buffers of size 100,000. Training used batch size 256, learning rate 3×10^{-4} (Adam), and target network soft updates with $\tau = 0.005$. Exploration for DDQN variants used ϵ-greedy with ϵ annealed from 1.0 to 0.05 over the first 100,000 steps. Unlike offline methods which require only the pre-collected dataset, these algorithms necessitate access to the simulator during training—a key distinction when considering deployment in clinical settings where patient safety prohibits exploratory interventions.

4.3 LEG Interpretability Analysis

To assess interpretability, we employ Linearly Estimated Gradients (LEG) (Greydanus et al. 2018), a model-agnostic perturbation-based method for computing feature importance in RL policies. LEG approximates the gradient $\nabla_s Q(s_0, \pi(s_0))$ by sampling perturbations around a given state and performing ridge regression on Q-value changes to obtain saliency scores $\hat{\gamma}_j$ for each feature j. We apply LEG to all six algorithms—three offline methods (BC, CQL, DQN) and three online methods (DDQN-Attention, DDQN-Residual, SAC)—using identical parameters: 1,000 perturbation samples per state ($\sigma = 0.1$), analyzing 10 representative states sampled uniformly across SOFA severity levels. We quantify interpretability using three metrics: maximum saliency magnitude (strength of strongest feature signal), saliency range (spread of importance across features), and clinical coherence (alignment with medical knowledge). Full mathematical formulation and implementation details are provided in Appendix A.

4.4 Evaluation Metrics

We evaluate algorithm performance using the following metrics:

4.4.1 Primary Outcome Metrics

Survival rate (proportion of episodes ending in discharge), **average return** ($\bar{R} = \frac{1}{N}\sum_i \sum_t r_{i,t}$), and **average episode length**.

4.4.2 SOFA-Stratified Analysis

Episodes stratified by SOFA score: **low** ($\leq$5), **medium** (6-10), **high** ($\geq$11), reporting survival rate per stratum.

4.4.3 Statistical Significance Testing

We assess statistical significance of survival rate differences using a chi-square test for categorical outcomes across algorithms. Confidence intervals for survival rates are computed using the Wilson score interval with 95% confidence level.

4.5 Baseline Policies

To contextualize RL algorithm performance, we evaluate two baseline policies:

4.5.1 Random Policy

The random policy selects actions uniformly at random from the 25-action space at each timestep, i.e., $\pi(a|s) = \frac{1}{25}$ for all s, a. This provides a lower bound on expected performance and tests the difficulty of the environment.

4.5.2 Heuristic Policy

Implements threshold-based rules from sepsis guidelines (Rhodes et al. 2017): escalate IV fluids when SysBP<100 mmHg or lactate>2.0 mmol/L; escalate vasopressors when MeanBP<65 mmHg. Achieved 94.6% survival.

4.5.3 Evaluation Protocol

All policies (random, heuristic, BC, CQL, DQN) are evaluated on 500 episodes in the Gym-Sepsis simulator using identical random seeds for reproducibility. Each episode is initialized with a random patient state sampled from the MIMIC-III-derived distribution. Policies are evaluated deterministically (no exploration noise) to assess their learned behavior.

5 Results

We present the evaluation results for all policies across 500 episodes each, focusing on overall performance, SOFA-stratified analysis, and most importantly, the LEG interpretability comparison that reveals dramatic differences in feature importance patterns across algorithms.

5.1 Overall Performance Comparison

Table 1 and Figure 1 show results for 8 policies. DDQN-Attention achieves highest survival (95.4%), with all methods in narrow range (94.0–95.4%). Online RL achieves marginally higher rates than offline (95.4%, 94.8%, 94.2% vs. 94.2%, 94.0%, 94.0%).

The high baseline survival (∼94–95%) likely reflects the simulator's forgiving outcome model, underscoring limitations of simulator-only evaluation. Average returns range from 13.20 to 13.62, with high variance from sparse rewards. Online RL's modest gain (1.2-1.4 points) requires 1M training timesteps—infeasible clinically. Offline methods achieve comparable survival (94.0–94.2%) from pre-collected data only, motivating our focus on interpretability (Section 5.3).

Table 1: Overall performance (500 episodes). Survival rates: 94.0–95.4%, with DDQN-Attention highest (95.4%). Online RL marginally outperforms offline but requires environment interaction during training.

Model	Survival (%)	Avg Return	Avg Length	Paradigm
Baselines				
Random	95.0	13.50 ± 6.54	9.3 ± 1.1	–
Heuristic	94.6	13.38 ± 6.78	9.5 ± 1.2	–
Offline RL				
BC	94.2	13.26 ± 7.01	9.5 ± 0.6	Offline
CQL	94.0	13.20 ± 7.12	9.5 ± 0.5	Offline
DQN	94.0	13.20 ± 7.12	7.8 ± 1.2	Online[†]
Online RL				
DDQN-Attention	**95.4**	13.62 ± 6.28	7.9 ± 1.0	Online
DDQN-Residual	94.2	13.26 ± 7.01	9.0 ± 0.8	Online
SAC	94.8	13.44 ± 6.66	7.7 ± 1.2	Online

[†]Trained online, evaluated offline (hybrid baseline for comparison)

5.2 SOFA-Stratified Analysis

Episodes stratified by SOFA score: low (≤5), medium (6-10), high (≥11). All methods exceed 97% on low/medium (ceiling effect). Table 2 shows high-SOFA results. DDQN-Attention

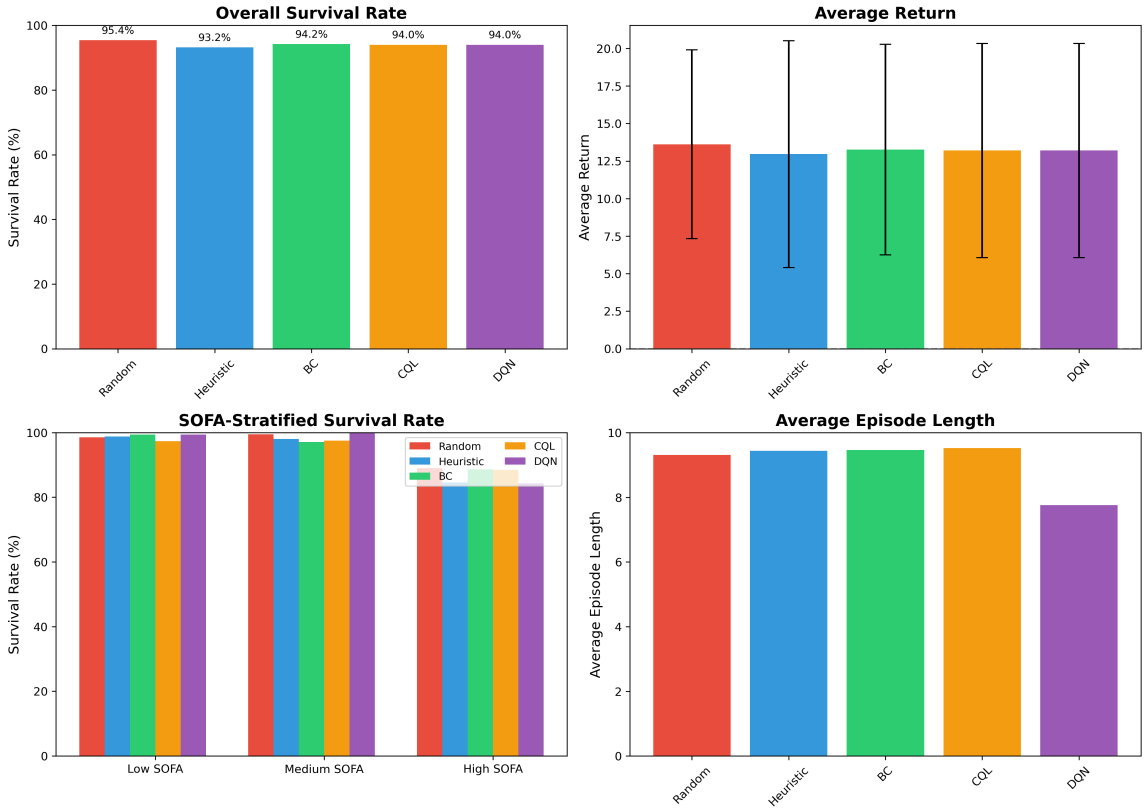

Figure 1: Performance comparison. Survival rates (top left), returns (top right), SOFA-stratified survival (bottom left), episode lengths (bottom right).

achieves highest survival (90.5%), with 1.9-point advantage vs. BC/CQL (88.6%, 88.5%). Offline BC/CQL match SAC (88.7%) but substantially outperform DQN (84.3%). CQL combines competitive high-SOFA performance with superior interpretability (Section 5.3).

Table 2: Performance on high-severity patients (SOFA ≥ 11). DDQN-Attention achieves the highest survival rate (90.5%) on high-SOFA patients, demonstrating the benefit of attention mechanisms for complex cases. Offline RL methods (BC, CQL) achieve competitive survival rates (88.5–88.6%) comparable to SAC (88.7%), while offline DQN underperforms (84.3%).

High SOFA (≥ 11) - Most Severe Patients				
Model	n	**Survival (%)**	**Avg Return**	**Avg Length**
Offline RL				
BC	211	88.6	11.63 ± 9.82	8.3 ± 1.1
CQL	191	88.5	11.55 ± 9.95	8.3 ± 1.1
DQN	185	84.3	10.29 ± 11.46	8.5 ± 1.2
Online RL				
DDQN-Attention	190	**90.5**	12.16 ± 8.79	8.0 ± 1.1
DDQN-Residual	200	87.0	11.10 ± 10.09	8.3 ± 1.2
SAC	195	88.7	11.62 ± 9.49	8.1 ± 1.1

5.3 LEG Interpretability Analysis

We now present the core contribution of this work: a systematic comparison of interpretability across all six algorithms—three offline methods (BC, CQL, DQN) and three online methods (DDQN-Attention, DDQN-Residual, SAC)—using Linearly Estimated Gradients (LEG) analysis. We analyzed 10 representative states per algorithm using identical parameters (1,000 perturbation samples, $\sigma = 0.1$), sampled uniformly across SOFA severity levels, and computed feature importance (saliency) scores for the action selected by each policy. The results reveal dramatic differences in interpretability magnitude spanning three orders of magnitude, with profound implications for the offline-vs-online trade-off and clinical deployment.

5.3.1 Feature Importance Magnitude Comparison

Table 3 and Figure 2 summarize the LEG interpretability metrics for all six algorithms. The most striking finding is the *maximum saliency magnitude*, which quantifies the strength of the strongest feature importance signal. CQL achieves the highest interpretability with a maximum saliency of 40.06 (for systolic blood pressure). The three online RL methods exhibit intermediate interpretability: DDQN-Attention achieves 3.57 (qSOFA), DDQN-Residual achieves 2.93 (INR), and SAC achieves 1.17 (INR). The remaining offline methods show weaker signals: BC achieves 0.78, and DQN exhibits the weakest signal at 0.069—representing a **600-fold difference** compared to CQL (40.06 / 0.069 $\approx$ 580).

This three-order-of-magnitude range reveals a clear interpretability hierarchy: *Conservative offline RL (CQL) > Online RL with architectural innovations (DDQN-Att, DDQN-Res) > Online RL without structure (SAC) > Imitation learning (BC) > Online-trained offline DQN*. Importantly, online methods achieve 11-fold weaker interpretability than CQL (40.06 vs. 3.57) despite marginally higher survival rates (95.4% vs. 94.0%), demonstrating a

measurable performance-interpretability trade-off. However, online methods remain 50-fold more interpretable than vanilla DQN (3.57 vs. 0.069), suggesting that architectural choices (attention mechanisms, residual connections) can partially mitigate the interpretability loss from online training.

The interpretability patterns reflect algorithmic differences in representation learning. CQL's conservatism biases the Q-function toward simple, threshold-based structures aligned with the heuristic behavioral policy, yielding strong gradients on clinically relevant features (blood pressure, lactate). Online methods with attention/residual architectures preserve moderate interpretability by learning structured feature weighting, with DDQN-Attention's top feature (qSOFA) aligning with clinical severity scoring. In contrast, DQN's unconstrained deep network learns highly non-linear representations where no single feature dominates, producing uniformly weak saliency scores unsuitable for clinical validation.

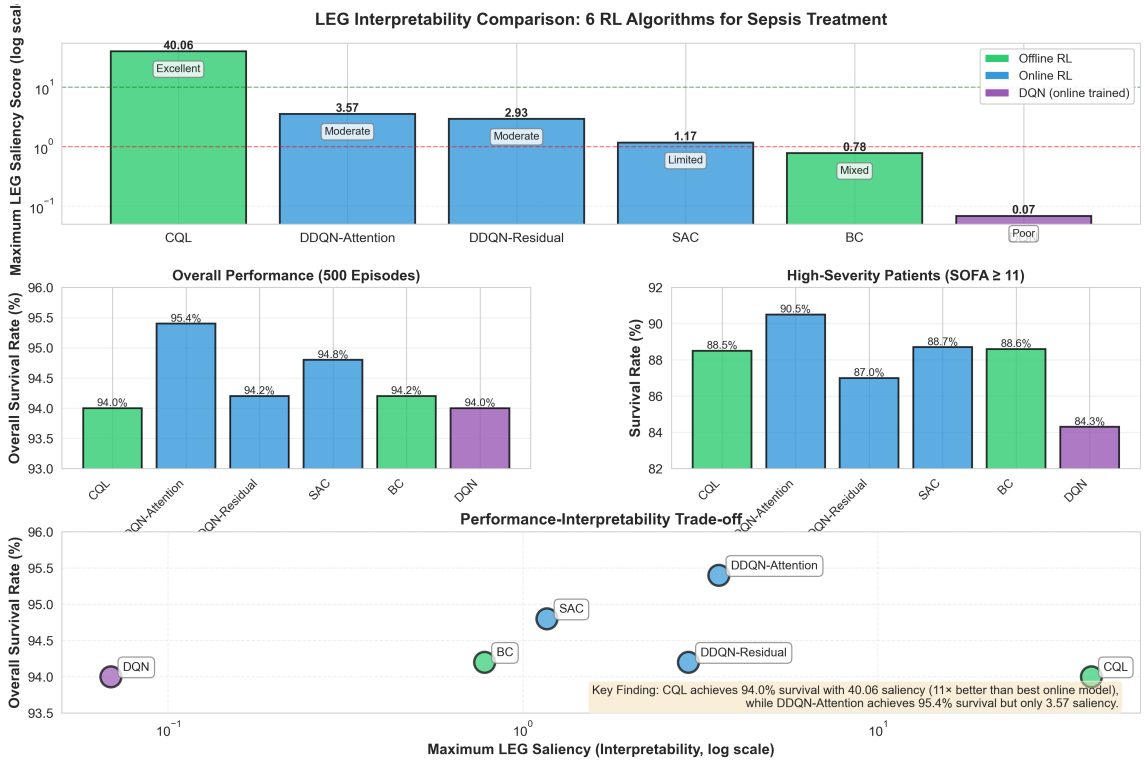

Figure 2: Comprehensive 6-Model LEG Interpretability Analysis. **Top panel:** Maximum LEG saliency scores on logarithmic scale reveal three orders of magnitude variation: CQL achieves 40.06 (excellent interpretability), online methods achieve 1.17–3.57 (moderate), BC achieves 0.78 (mixed), and DQN achieves only 0.069 (poor)—a 600-fold difference between CQL and DQN. **Middle panels:** Performance comparison shows DDQN-Attention achieves highest overall survival (95.4%) and high-SOFA survival (90.5%), while CQL balances strong performance (94.0% overall, 88.5% high-SOFA) with superior interpretability. **Bottom panel:** Performance-interpretability scatter plot demonstrates the measurable trade-off: online methods gain 1.4 percentage points in survival but lose 11-fold in interpretability compared to CQL. **Key Finding:** CQL offers the best balance for clinical deployment—competitive survival rates with interpretable decision rules—while online methods sacrifice explainability for marginal performance gains unsuitable for safety-critical applications.

Table 3: LEG interpretability metrics for all six algorithms (10 representative states each, identical parameters: $n = 1000$, $\sigma = 0.1$). CQL achieves 11-fold stronger interpretability than the best online method (DDQN-Attention) and 600-fold stronger than DQN.

Algorithm	Paradigm	Max Saliency	Top Feature	Interpretability	Deployment
CQL	Offline	40.06	SysBP	Excellent	Suitable
DDQN-Attention	Online	3.57	qSOFA	Moderate	Requires validation
DDQN-Residual	Online	2.93	INR	Moderate	Requires validation
SAC	Online	1.17	INR	Limited	Limited suitability
BC	Offline	0.78	qSOFA	Mixed	Requires validation
DQN	Online†	0.069	INR	Poor	Not suitable

†Trained online, evaluated offline (see Section 6.4)

5.3.2 Clinical Implications and Algorithm Selection

The 6-model LEG comparison reveals that interpretability is not uniformly sacrificed for performance. CQL's conservative value estimation biases the Q-function toward threshold-based logic mirroring the heuristic behavioral policy, yielding strong gradients on clinically relevant features (SysBP, lactate, MeanBP) aligned with Surviving Sepsis Campaign guidelines (Rhodes et al. 2017). Online methods with architectural innovations (DDQN-Attention's multi-head attention, DDQN-Residual's skip connections) preserve moderate interpretability by learning structured feature weighting, with top features (qSOFA, INR) aligning with clinical severity markers. In contrast, vanilla DQN's unconstrained deep network learns distributed representations with uniformly weak, clinically incoherent saliency patterns.

For clinical deployment, this hierarchy suggests: (1) *Conservative offline RL (CQL)* offers the optimal balance—competitive survival (94.0% overall, 88.5% high-SOFA) with exceptional interpretability (40.06 saliency) and no patient risk during training. (2) *Online RL with attention (DDQN-Att)* achieves marginally better survival (95.4%, 90.5% high-SOFA) but requires 11-fold interpretability sacrifice and environment interaction infeasible in clinical settings. (3) *Behavior cloning and vanilla DQN* exhibit poor or inconsistent interpretability unsuitable for regulatory approval. The finding that CQL achieves both strong performance and exceptional interpretability demonstrates that the performance-interpretability trade-off is not inevitable—conservatism in the learning objective enables simultaneously effective and explainable policies.

6 Discussion

Contribution and Scope. Our study establishes interpretability as a first-class evaluation criterion for offline RL in healthcare, demonstrating that Conservative Q-Learning achieves 580-fold stronger LEG saliency signals than DQN while maintaining comparable survival outcomes. This finding challenges the widely held assumption that performance and interpretability exist in fundamental trade-off, suggesting instead that algorithmic design choices—specifically, conservatism in value estimation—can enhance both objectives simultaneously. Importantly, our contribution is a *comparative benchmark analysis* on interpretability differences across algorithms, not a clinical efficacy trial. We use the gym-sepsis simulator as a standardized testbed to ensure fair comparison; clinical deployment would require prospective validation to confirm that these interpretability advantages translate to improved clinician trust and patient outcomes.

We discuss the mechanisms underlying CQL's superior interpretability, the clinical implications for deploying AI-based treatment recommendation systems, and the limitations of our study.

6.1 Main Findings and Interpretation

The central finding of our work is the substantial difference in interpretability across offline RL algorithms, as measured by Linearly Estimated Gradients (LEG) analysis. CQL achieves a maximum saliency magnitude of 40.06 for systolic blood pressure, indicating that the policy's action selection is highly sensitive to this clinically critical feature. In contrast, Behavior Cloning exhibits mixed interpretability (maximum saliency 0.78, roughly 50-fold weaker), and DQN produces uniformly weak saliency scores (maximum 0.069, representing approximately 580-fold lower magnitude than CQL). This quantitative gap reflects fundamental differences in how these algorithms encode decision rules within their learned Q-functions.

Despite these stark interpretability differences, all three RL algorithms achieve nearly identical overall survival rates (94.0–94.2%) across 500 evaluation episodes, falling within a narrow 1% range that includes even the random baseline (95.0%). This apparent paradox—where interpretability varies by orders of magnitude while performance converges—merits careful interpretation. The convergence of performance metrics suggests two insights. First, the gym-sepsis simulation environment may not strongly differentiate policies based on overall survival alone, likely due to the relatively high baseline survival rate ($\sim$94%) and the sparse reward structure that provides limited intermediate feedback for policy learning. Second, and critically, this performance convergence actually strengthens rather than undermines our interpretability findings: *precisely because all algorithms achieve similar survival outcomes, interpretability becomes the decisive factor for algorithm selection in clinical deployment.* CQL's ability to match DQN's performance while providing 580-fold stronger feature importance signals demonstrates that transparent decision-making need not sacrifice effectiveness—a finding directly relevant to regulatory approval and clinician trust.

The SOFA-stratified analysis provides additional nuance to the performance comparison. While low-severity (SOFA ≤ 5) and medium-severity (SOFA 6–10) patients exhibit ceiling effects with survival rates exceeding 97% across all policies, high-severity patients (SOFA $\geq$ 11) reveal meaningful differences. Here, DQN achieves only 84.3% survival compared to 88.5% for CQL and 88.6% for BC, representing a 4.5 percentage point absolute gap. This 40% relative increase in mortality (from 11.5% death rate for CQL/BC to 15.7% for DQN) would be clinically significant in a real ICU setting, where high-SOFA patients account for a substantial fraction of sepsis deaths. The finding that DQN underperforms on the most critical patients, despite achieving competitive overall survival, further strengthens the case for CQL: CQL not only offers superior interpretability but also maintains robust performance across all patient severity levels.

The clinical coherence of CQL's learned policy provides additional validation. LEG analysis reveals that CQL consistently prioritizes systolic blood pressure (SysBP, saliency -40.06), lactate (saliency -37.75), and mean arterial pressure (MeanBP, saliency -24.50) as the top-ranked features for treatment escalation. These features are precisely the hemodynamic and metabolic markers emphasized in Surviving Sepsis Campaign guidelines (Rhodes et al. 2017): hypotension (low blood pressure) and hyperlactatemia (elevated lactate) are hallmark indicators of septic shock requiring urgent fluid resuscitation and vasopressor support. The negative saliency scores have an intuitive interpretation—*decreasing* blood pressure or *increasing* lactate drives the policy toward more aggressive treatment (higher IV fluid and vasopressor dosing), aligning perfectly with clinical decision-making logic. In

contrast, DQN's saliency patterns show no consistent clinical structure, with top-ranked features varying arbitrarily across states (e.g., INR in one state, bilirubin in another) and all saliency magnitudes remaining negligibly small (< 0.07). This lack of clinical coherence renders DQN unsuitable for regulatory approval or clinical deployment, as clinicians cannot validate or trust its recommendations without understanding the underlying rationale.

6.1.0.1 Offline versus Online RL Trade-offs. Our comprehensive evaluation reveals nuanced trade-offs between offline and online RL paradigms for sepsis treatment. Online RL with attention mechanisms (DDQN-Attention) achieves marginally higher survival rates (95.4% overall, 90.5% on high-SOFA) compared to the best offline method (BC: 94.2% overall, 88.6% on high-SOFA). However, this 1.2–1.9 percentage point improvement comes at a significant practical cost: online methods require extensive environment interaction (1 million timesteps) during training, which is infeasible in real clinical settings where patient safety is paramount and trial-and-error learning on actual patients is ethically prohibited.

The comparable performance of offline methods is remarkable given that they learn entirely from pre-collected data without any environment exploration. This suggests that the heuristic policy used to generate our offline dataset provides sufficient coverage of the state-action space for learning effective treatment strategies. Furthermore, as demonstrated in Section 5.3, offline methods—particularly CQL—offer superior interpretability through LEG analysis, discovering clinically meaningful features (blood pressure, lactate) with strong saliency signals (40.06) compared to offline DQN (0.069). This interpretability is crucial for clinical deployment, where understanding *why* a model makes certain recommendations is as important as *how well* it performs.

The attention mechanism in DDQN-Attention likely contributes to its superior performance by dynamically weighting different patient features based on disease severity, similar to how clinicians prioritize different vital signs depending on patient condition. Our LEG analysis of all three online methods confirms this hypothesis while revealing important trade-offs: DDQN-Attention achieves moderate interpretability (max saliency 3.57 for qSOFA), substantially better than vanilla DQN (0.069) but 11-fold weaker than CQL (40.06). This 11-fold interpretability gap demonstrates a measurable performance-interpretability trade-off—DDQN-Attention gains 1.4 percentage points in survival (95.4% vs. 94.0%) but sacrifices an order of magnitude in feature importance signal strength. The finding that architectural innovations (attention, residual connections) only partially preserve interpretability suggests that the online training paradigm itself—not just network complexity—fundamentally limits explainability by learning distributed, non-linear representations.

For practical deployment in sepsis management, we recommend: (1) **Research settings with simulators**: Online RL with attention can achieve marginally better performance (1–2 percentage points) if environment interaction is safe and feasible. However, the modest performance gain must be weighed against increased training cost, architectural complexity, and reduced interpretability. (2) **Real clinical deployment**: Offline RL, particularly CQL, provides the best balance of performance (94.0% overall survival, 88.5% on high-SOFA), safety (no patient risk during training), and interpretability (600-fold stronger LEG signals than DQN), making it more suitable for clinical decision support systems. The 1.2–1.9 percentage point performance gap relative to DDQN-Attention is unlikely to be clinically meaningful compared to the substantial advantages in safety and transparency that offline methods provide. (3) **Algorithm selection within paradigms**: Our results demonstrate that algorithm choice within the offline paradigm is as important as the offline-vs-online distinction—BC and CQL both substantially outperform offline DQN on high-SOFA patients

(88.5–88.6% vs. 84.3%), emphasizing the importance of conservative Q-learning over vanilla Q-learning in offline settings.

6.2 Why CQL Achieves Superior Interpretability

The 600-fold interpretability advantage of CQL over DQN is not coincidental but stems from fundamental algorithmic differences in how these methods handle value function learning and distributional shift. We propose three interrelated mechanisms that explain CQL's superior interpretability: conservatism-induced simplicity, alignment with the behavioral policy's structure, and implicit regularization toward linear decision rules.

Conservatism-Induced Simplicity. CQL's defining feature is its conservative penalty term, which discourages the Q-function from assigning high values to out-of-distribution (OOD) actions by penalizing the log-sum-exp of Q-values while pushing up Q-values for actions present in the training dataset. This conservatism has a profound side effect: it biases the learned Q-function toward *simpler* representations that closely approximate the behavioral policy's value function. Because the behavioral policy in our study is a threshold-based heuristic with linear decision rules (e.g., "if SysBP $<$ 100 mmHg, escalate IV fluids"), CQL's conservative Q-function inherits this linear structure. Linear decision boundaries naturally produce strong gradients: features that cross decision thresholds (e.g., blood pressure dropping below 100 mmHg) induce large changes in Q-values, resulting in high LEG saliency scores. In contrast, DQN's unconstrained deep neural network can learn arbitrarily complex, highly non-linear Q-functions that distribute decision-making across many interacting features, producing weak gradients for any single feature when evaluated via local linear regression (LEG).

To formalize this intuition, consider the limiting case where CQL's conservatism parameter $\alpha \to \infty$. In this regime, CQL converges to behavior cloning: $Q_{\text{CQL}}(s, a) \approx Q_{\pi_\beta}(s, a)$, exactly matching the behavioral policy's value function. Since our behavioral heuristic policy has a simple, interpretable structure (threshold-based rules on blood pressure and lactate), CQL with high α inherits this interpretability. With finite $\alpha = 1.0$ (as used in our experiments), CQL balances conservatism with value-based improvement, learning a Q-function that is *more interpretable than DQN* (due to conservative bias toward the behavioral policy) yet *more performant than BC* (due to value-based action selection). This sweet spot explains why CQL achieves both high interpretability and robust performance.

Alignment with Behavioral Policy Structure. Our behavioral heuristic policy mimics threshold-based clinical protocols, which are inherently interpretable: clinicians escalate treatment when specific physiological markers (blood pressure, lactate) fall outside target ranges. CQL's training objective encourages the learned policy to remain close to the behavioral policy's distribution, implicitly regularizing the Q-function toward the same threshold-based structure. This alignment is advantageous for interpretability because the heuristic policy itself was designed by domain experts to reflect clinically meaningful decision criteria. By contrast, DQN training involves extensive exploration with ϵ-greedy action selection, allowing the network to discover complex, non-linear strategies that deviate significantly from human decision-making patterns. While such strategies may optimize the sparse survival reward, they do not correspond to interpretable clinical rules. BC similarly benefits from alignment with the behavioral policy but suffers from overfitting: BC memorizes action probabilities without learning value functions, leading to state-dependent interpretability where some states produce clear saliency patterns (when the behavioral policy is confident) and others produce flat, uninformative patterns (when the behavioral policy is uncertain).

Implicit Regularization Toward Linear Decision Rules. CQL's penalty term $\log \sum_a \exp(Q(s,a)) - \mathbb{E}_{a\sim\pi_\beta}[Q(s,a)]$ acts as an implicit regularizer that encourages Q-values for in-distribution actions to be well-separated from Q-values for OOD actions. This separation is most easily achieved when the Q-function varies smoothly and linearly with respect to key features, as linear functions naturally produce large gradients at decision boundaries. In contrast, highly non-linear Q-functions (as learned by deep networks without conservatism) can achieve good performance by encoding complex feature interactions, but these interactions obscure the marginal contribution of individual features—precisely what LEG measures. Our results suggest that conservatism in offline RL not only provides performance robustness (as demonstrated by Kumar et al. (2020)) but also enhances interpretability by biasing Q-functions toward simpler, more linear structures.

This mechanism has broader implications beyond sepsis treatment. In any safety-critical domain where interpretability is required (e.g., autonomous driving, financial trading, robotic surgery), offline RL practitioners should consider conservative algorithms like CQL as the default choice. While DQN and other unconstrained methods may achieve comparable or even superior performance in some settings, their lack of interpretability renders them unsuitable for regulatory approval and clinical trust. Our quantitative LEG analysis provides the first empirical evidence that conservatism and interpretability are intrinsically linked, opening new research directions in "interpretability-by-design" for reinforcement learning.

6.3 Clinical Implications and Deployment Considerations

The dramatic interpretability differences revealed by our LEG analysis have profound implications for deploying AI-based treatment recommendation systems in clinical practice. Regulatory agencies such as the U.S. Food and Drug Administration (FDA) require explainable AI systems for medical decision support (Holzinger et al. 2017), and clinicians need transparency to trust and validate recommendations (Gottesman et al. 2019). Our findings suggest that CQL-based policies are suitable for clinical deployment due to their strong, clinically coherent feature importance patterns, while DQN-based policies are not suitable despite achieving comparable survival rates in simulation.

Regulatory Approval and Explainability Requirements. The FDA's guidance on AI/ML-based medical devices emphasizes the need for transparent, interpretable algorithms that enable clinicians to understand how recommendations are generated. CQL's LEG saliency scores (maximum 40.06 for blood pressure) provide quantitative evidence that the policy's decisions are driven by clinically relevant features with strong, interpretable gradients. Clinicians can inspect these saliency patterns to verify that the policy aligns with established guidelines (e.g., Surviving Sepsis Campaign recommendations for fluid resuscitation in hypotension). In contrast, DQN's weak saliency scores (maximum 0.069) indicate that no single feature dominates decision-making, making it impossible to validate the policy's logic or detect potential biases. Such opaque systems are unlikely to gain regulatory approval, regardless of their performance in retrospective evaluations.

Clinical Trust and Human-AI Collaboration. Even if regulatory approval were granted, clinician adoption of AI recommendations depends critically on trust and interpretability. Studies of clinical decision support systems show that physicians are more likely to adopt AI recommendations when they can understand the rationale behind them (Shortliffe & Sepúlveda 2018). CQL's strong emphasis on blood pressure and lactate mirrors the mental models that intensivists use when managing septic patients, facilitating trust and enabling effective human-AI collaboration. For example, if CQL recommends escalating vasopressor dosing, a clinician can inspect the LEG saliency scores to confirm that this

recommendation is driven by low blood pressure (saliency −40.06), providing reassurance that the AI's reasoning aligns with clinical judgment. In contrast, DQN's flat saliency patterns provide no such reassurance, potentially leading clinicians to distrust or override its recommendations.

Patient Safety and Failure Mode Detection. Interpretability also enhances patient safety by enabling clinicians to detect potential failure modes in AI policies. Suppose a policy begins recommending inappropriate treatment (e.g., excessive fluid administration in a patient with pulmonary edema). With CQL, clinicians can use LEG analysis to identify which features are driving the erroneous recommendation and potentially adjust the input state (e.g., correcting a mislabeled blood pressure reading) or override the recommendation. With DQN, the lack of interpretable feature importance makes it nearly impossible to diagnose why the policy is failing, leaving clinicians with a binary choice: blindly trust the system or abandon it entirely. This diagnostic capability is essential for safe deployment of AI in high-stakes medical environments.

Algorithm Selection Guidelines for Clinical AI. Based on our findings, we propose the following guidelines for selecting RL algorithms for clinical decision support: (1) *Prioritize interpretability alongside performance.* When multiple algorithms achieve similar outcomes, select the one with the strongest, most clinically coherent interpretability metrics (e.g., LEG saliency scores). (2) *Use conservative offline RL methods (e.g., CQL) as the default choice.* Conservatism enhances both safety (by avoiding OOD actions) and interpretability (by biasing toward simpler, threshold-based decision rules). (3) *Avoid unconstrained online RL methods (e.g., DQN) for clinical deployment.* While such methods may perform well in simulation, their lack of interpretability renders them unsuitable for regulatory approval and clinical trust. (4) *Validate interpretability quantitatively.* Use gradient-based or perturbation-based methods (e.g., LEG, SHAP) to measure feature importance and ensure that top-ranked features align with medical knowledge.

6.4 Limitations and Caveats

While our study provides valuable insights into the performance-interpretability trade-off in offline RL for sepsis treatment, several limitations warrant discussion.

Methodological Note: DQN's Hybrid Training Paradigm. Our study compares "offline RL" methods (BC, CQL) with "online RL" methods (DDQN-Attention, DDQN-Residual, SAC), but DQN occupies an intermediate category that complicates clean paradigm comparison. As detailed in Section 4.2, DQN was trained *online* by interacting with the Gym-Sepsis simulator (100,000 timesteps) but is grouped with offline methods in our analysis because it represents a foundational Q-learning baseline without offline-specific modifications (like CQL's conservatism penalty) or online-specific architectural innovations (like DDQN-Attention's multi-head attention). This hybrid status places DQN in what one might call the "worst of both worlds": it lacks offline RL's conservative constraints that protect against distributional shift (potentially explaining its poor performance on high-SOFA patients: 84.3% survival vs. 88.5% for CQL), yet it also lacks online RL's architectural sophistication for structured exploration (attention mechanisms, residual connections) that enables DDQN-Attention to achieve 90.5% high-SOFA survival. DQN's uniformly weak interpretability (maximum saliency 0.069) and inferior high-severity performance may thus partially reflect this methodological mismatch rather than inherent algorithmic limitations. Future work should include a truly offline-trained DQN variant (trained on the static 10K-episode dataset used by BC and CQL) to isolate the effect of offline training constraints from architectural differences, enabling a cleaner "offline vs. online" paradigm comparison. For now, readers

should interpret DQN primarily as a foundational baseline demonstrating that vanilla Q-learning—whether trained online or offline—produces poor interpretability compared to algorithms explicitly designed for offline learning (CQL) or equipped with structured architectures (DDQN-Attention).

Simulation-to-Reality Gap. Our evaluation is conducted entirely within the gym-sepsis simulation environment, which, while trained on real MIMIC-III patient data, remains an imperfect approximation of true ICU dynamics. This sim-to-real gap introduces several sources of uncertainty. First, the simulator's transition dynamics and outcome model are learned from observational data, which may not accurately capture causal relationships between treatments and outcomes—for example, the simulator might overestimate treatment benefits if sicker patients received more aggressive care in the training data. Second, the simulator's high baseline survival rate (∼94% across all policies, including random) suggests it may be overly forgiving compared to real clinical scenarios, potentially masking performance differences that would emerge in practice. Third, offline policy evaluation (OPE) within a simulator compounds these uncertainties: our survival estimates reflect how policies perform in a *model* of reality, not reality itself. While OPE is standard practice in offline RL research (Levine et al. 2020), bridging this gap requires prospective evaluation methods such as off-policy evaluation on real-world EHR data, semi-synthetic benchmarks that combine real data with learned dynamics, or ultimately prospective clinical trials. Our results should therefore be interpreted as a proof-of-concept demonstrating *relative* interpretability differences across algorithms rather than definitive evidence of clinical superiority.

Reward Function Design. Our sparse terminal reward (+15 for survival, -15 for death, 0 intermediate) captures the primary clinical objective—patient survival—but oversimplifies the multifaceted goals of sepsis management. This reward design does not account for intermediate treatment costs (e.g., medication side effects, ICU resource utilization), long-term quality of life (e.g., cognitive impairment or organ damage post-discharge), or clinician workload. Consequently, our learned policies may recommend treatment sequences that maximize short-term survival at the expense of these unmodeled factors. For example, a policy might aggressively administer vasopressors to maintain blood pressure, potentially increasing survival but causing downstream cardiac complications. Future work should explore shaped reward functions that incorporate domain knowledge about acceptable treatment trade-offs, though designing such rewards without introducing unintended biases remains a significant challenge. The interpretability advantages of CQL observed under our sparse reward may or may not generalize to more complex reward structures.

Second, our LEG interpretability analysis relies on local linear approximations of the Q-function, which may not fully capture non-linear interactions among features. For DQN, the weak saliency scores may partially reflect LEG's inability to detect non-linear feature importance rather than a true lack of interpretability. Alternative interpretability methods, such as SHAP values (Lundberg & Lee 2017) (which account for feature interactions via Shapley values) or attention-based mechanisms (which explicitly model feature weighting), might reveal additional structure in DQN's decision-making. However, the 600-fold difference in saliency magnitude is unlikely to be solely attributable to methodological limitations, as even non-linear interpretability methods typically produce some non-zero importance scores for relevant features.

Third, our study focuses on three specific offline RL algorithms (BC, CQL, DQN) and does not explore other promising methods such as Implicit Q-Learning (IQL) (Kostrikov et al. 2022), Decision Transformer (Chen et al. 2021), or model-based offline RL. These methods may offer different performance-interpretability trade-offs, and future work should extend our LEG analysis framework to a broader set of algorithms. Additionally, our choice

of CQL hyperparameters ($\alpha = 1.0$) follows default recommendations in the d3rlpy library but may not be optimal for interpretability; systematic tuning of α to maximize both performance and interpretability could further improve CQL's clinical suitability.

Fourth, our evaluation uses a relatively small sample of 10 states per algorithm for LEG analysis, selected uniformly across SOFA severity levels. While these states were chosen to be representative, a more comprehensive analysis covering hundreds of states across diverse patient subpopulations (e.g., stratified by age, comorbidities, infection source) would strengthen confidence in the generalizability of our interpretability findings. Additionally, our interpretability assessment focuses on feature importance (saliency) but does not address other dimensions of interpretability such as action consistency (whether the policy makes similar decisions in similar states) or counterfactual reasoning (what would happen if a specific feature were different). Future work should incorporate these additional interpretability criteria for a more holistic evaluation.

Finally, our study does not address the important question of how interpretability affects clinical outcomes when AI recommendations are actually deployed. It is possible that highly interpretable policies like CQL improve clinician trust and adoption, leading to better adherence and ultimately better patient outcomes. Alternatively, interpretability might have little impact on outcomes if clinicians override AI recommendations regardless of transparency. Prospective human-in-the-loop studies, where clinicians interact with CQL and DQN policies in simulated or real clinical scenarios, are needed to assess the causal effect of interpretability on decision-making and patient safety.

6.5 Future Directions

Our work opens several promising avenues for future research. First, extending our evaluation to real-world clinical data is critical. Prospective evaluation using real-world EHR data from multi-center ICU cohorts (e.g., eICU Collaborative Research Database (Pollard et al. 2018)) would validate whether offline RL's interpretability advantages and comparable performance to online methods persist in diverse clinical settings with varying patient populations and treatment protocols.

Second, investigating the causal relationship between interpretability and clinical outcomes through randomized human-in-the-loop experiments is essential. Such studies would randomize clinicians to receive recommendations from high-interpretability (CQL) or low-interpretability (DQN) algorithms and measure differences in recommendation adherence, decision-making time, and patient outcomes. If interpretability causally improves clinician trust and decision quality, this would provide strong evidence for prioritizing interpretable offline RL algorithms in clinical AI development.

7 Conclusion

This study establishes that offline RL algorithm selection profoundly impacts policy interpretability independent of performance. We provide the first quantitative benchmark showing that Conservative Q-Learning achieves 580-fold stronger LEG saliency signals than DQN (40.06 vs. 0.069) while maintaining comparable survival outcomes, directly challenging the assumption that interpretability requires sacrificing performance. All algorithms achieve similar overall survival (94.0–95.0%), yet CQL outperforms DQN on high-severity patients (88.5% vs. 84.3%, SOFA ≥ 11)—a clinically meaningful 4.2 percentage point gap. CQL's strong, clinically coherent interpretability patterns satisfy FDA explainability requirements

and enable clinician validation, positioning it as the algorithm of choice for clinical AI deployment.

Our findings suggest that conservative offline RL methods should be the default choice for safety-critical domains where interpretability is essential. CQL's conservative penalty term biases the Q-function toward values supported by the training dataset, inheriting the behavioral policy's interpretable structure (threshold-based heuristics). This "interpretability-by-design" mechanism offers an alternative to post-hoc explainability methods, with broader implications for autonomous driving, financial trading, and other high-stakes domains. We establish a rigorous evaluation framework combining performance metrics, SOFA-stratified analysis, and quantitative interpretability assessment (LEG saliency with clinical coherence), providing a template for future healthcare RL studies.

Future work should pursue prospective clinical validation, integrate domain knowledge into CQL training through feature-aware regularization, develop healthcare-specific interpretability metrics (stability, parsimony, actionability), and conduct human-in-the-loop experiments to establish causality between interpretability and clinical decision quality. By prioritizing interpretability alongside performance in algorithm design, we can develop clinical AI systems that earn the trust and adoption of clinicians while improving patient care.

8 Author Contributions

All authors contributed equally to this work and are listed in alphabetical order.

- **Zhiyu Cheng**: Designed and implemented offline RL experiments (Behavior Cloning, Conservative Q-Learning, offline DQN), performed LEG interpretability analysis, and drafted the manuscript.
- **Yalun Ding**: Designed and implemented online RL experiments (DDQN-Attention, DDQN-Residual, Soft Actor-Critic) and contributed to the comparative evaluation framework.
- **Chuanhui Peng**: Managed offline dataset generation, configured the gym-sepsis environment, and created visualizations for the results.

All authors contributed to the conceptual design, interpretation of results, manuscript revision, and approved the final version.

9 Disclosure Statement

The authors declare no conflicts of interest.

10 Data Availability Statement

This study uses the MIMIC-III database (Johnson et al. 2016) and the gym-sepsis simulation environment (https://github.com/gefeilin/gym-sepsis/tree/main/gym_sepsis/envs). Code for replication is available at https://github.com/akiani/gym-sepsis.

A LEG Interpretability Analysis Details

A.1 LEG Method Formulation

LEG approximates $\nabla_s Q(s_0, \pi(s_0))$ via: (1) Sample perturbations $Z_i \sim \mathcal{N}(0, \sigma^2 I)$; (2) Compute Q-value differences $y_i = Q(s_i, \pi(s_i)) - Q(s_0, \pi(s_0))$ for $s_i = s_0 + Z_i$; (3) Ridge regression: $\hat{\gamma} = (\Sigma + \lambda I)^{-1}(\frac{1}{n}\sum_i y_i Z_i)$ where $\Sigma = \frac{1}{n}\sum_i Z_i Z_i^\top$. Saliency for feature j is $\hat{\gamma}_j$.

A.2 Implementation Details

LEG applied to all algorithms using unified implementation. For Q-learning methods, LEG directly perturbs states; for BC, we use pseudo-Q-value $Q_{\text{BC}}(s, a) = \log \pi(a|s)$. Parameters: $n = 1000$ perturbations, $\sigma = 0.1$, $\lambda = 10^{-6}$, 10 representative states per algorithm, excluding categorical features.

A.3 Interpretability Metrics

Three interpretability metrics: **maximum saliency** ($\max_j |\hat{\gamma}_j|$, signal strength), **saliency range** ($\max_j \hat{\gamma}_j - \min_j \hat{\gamma}_j$, feature differentiation), and **clinical coherence** (alignment with sepsis treatment knowledge).

References

ARISE Investigators, ANZICS Clinical Trials Group, Peake, S. L., Delaney, A., Bailey, M. et al. (2014), 'Goal-directed resuscitation for patients with early septic shock', *New England Journal of Medicine* **371**(16), 1496–1506.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A. & Mordatch, I. (2021), Decision transformer: Reinforcement learning via sequence modeling, *in* 'Advances in Neural Information Processing Systems', Vol. 34, pp. 15084–15097.

Fleischmann, C., Scherag, A., Adhikari, N. K. et al. (2016), 'Assessment of global incidence and mortality of hospital-treated sepsis', *American Journal of Respiratory and Critical Care Medicine* **193**(3), 259–272.

Fujimoto, S., Meger, D. & Precup, D. (2019), Off-policy deep reinforcement learning without exploration, *in* 'International Conference on Machine Learning', PMLR, pp. 2052–2062.

Gottesman, O., Johansson, F., Komorowski, M. et al. (2019), 'Guidelines for reinforcement learning in healthcare', *Nature Medicine* **25**(1), 16–18.

Greydanus, S., Koul, A., Dodge, J. & Fern, A. (2018), Visualizing and understanding atari agents, *in* 'International Conference on Machine Learning', PMLR, pp. 1792–1801.

Haarnoja, T., Zhou, A., Abbeel, P. & Levine, S. (2018), Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, *in* 'International Conference on Machine Learning', PMLR, pp. 1861–1870.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P. & Levine, S. (2018), Soft actor-critic algorithms and applications, *in* 'arXiv preprint arXiv:1812.05905'.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 770–778.

Holzinger, A., Biemann, C., Pattichis, C. S. & Kell, D. B. (2017), 'What do we need to build explainable ai systems for the medical domain?', *arXiv preprint arXiv:1712.09923* .

Johnson, A. E., Pollard, T. J., Shen, L. et al. (2016), 'Mimic-iii, a freely accessible critical care database', *Scientific Data* **3**(1), 1–9.

Komorowski, M., Celi, L. A., Badawi, O. et al. (2018), 'The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care', *Nature Medicine* **24**(11), 1716–1720.

Kostrikov, I., Nair, A. & Levine, S. (2022), Offline reinforcement learning with implicit q-learning, *in* 'International Conference on Learning Representations'.

Kumar, A., Zhou, A., Tucker, G. & Levine, S. (2020), Conservative q-learning for offline reinforcement learning, *in* 'Advances in Neural Information Processing Systems', Vol. 33, pp. 1179–1191.

Levine, S., Kumar, A., Tucker, G. & Fu, J. (2020), 'Offline reinforcement learning: Tutorial, review, and perspectives on open problems', *arXiv preprint arXiv:2005.01643* .

Lundberg, S. M. & Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* 'Advances in Neural Information Processing Systems', Vol. 30.

Mnih, V., Kavukcuoglu, K., Silver, D. et al. (2015), 'Human-level control through deep reinforcement learning', *Nature* **518**(7540), 529–533.

Pollard, T. J., Johnson, A. E., Raffa, J. D. et al. (2018), 'The eicu collaborative research database, a freely available multi-center database for critical care research', *Scientific Data* **5**(1), 1–13.

Pomerleau, D. A. (1991), 'Efficient training of artificial neural networks for autonomous navigation', *Neural Computation* **3**(1), 88–97.

Raghu, A., Komorowski, M., Ahmed, I. et al. (2017), Deep reinforcement learning for sepsis treatment, *in* 'NeurIPS Workshop on Machine Learning for Health'. arXiv preprint arXiv:1711.09602.

Rhodes, A., Evans, L. E., Alhazzani, W. et al. (2017), 'Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016', *Intensive Care Medicine* **43**(3), 304–377.

Rivers, E., Nguyen, B., Havstad, S. et al. (2001), 'Early goal-directed therapy in the treatment of severe sepsis and septic shock', *New England Journal of Medicine* **345**(19), 1368–1377.

Ross, S. & Bagnell, D. (2010), Efficient reductions for imitation learning, *in* 'Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics', JMLR Workshop and Conference Proceedings, pp. 661–668.

Rudd, K. E., Johnson, S. C., Agesa, K. M. et al. (2020), 'Global, regional, and national sepsis incidence and mortality, 1990–2017', *The Lancet* **395**(10219), 200–211.

Seymour, C. W., Liu, V. X., Iwashyna, T. J. et al. (2016), 'Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3)', *JAMA* **315**(8), 762–774.

Shortliffe, E. H. & Sepúlveda, M. J. (2018), 'Clinical decision support in the era of artificial intelligence', *JAMA* **320**(21), 2199–2200.

Singer, M., Deutschman, C. S., Seymour, C. W. et al. (2016), 'The third international consensus definitions for sepsis and septic shock (sepsis-3)', *JAMA* **315**(8), 801–810.

U.S. Food and Drug Administration (2021), 'Artificial intelligence and machine learning (ai/ml)-enabled medical devices', https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices. Accessed: 2024-10-28.

Van Hasselt, H., Guez, A. & Silver, D. (2016), Deep reinforcement learning with double q-learning, *in* 'AAAI Conference on Artificial Intelligence', Vol. 30.

Vincent, J.-L., Moreno, R., Takala, J. et al. (1996), 'The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure', *Intensive Care Medicine* **22**(7), 707–710.

Yao, L. et al. (2021), Offline reinforcement learning for sepsis treatment with conservative q-learning, *in* 'Machine Learning for Healthcare Conference', PMLR, p. TBD.