Dense Reward Discovery for Sparse Outcomes: Contextual Inverse Reinforcement Learning for Clinical Trajectories

Anonymous Authors[1]

Abstract

A persistent challenge in medical reinforcement learning (RL) is learning from long trajectories with mixed observability and sparse outcomes. Consider ICU treatment for septic shock: patients are hospitalized for days with continuous treatment decisions, yet the primary outcome—patient survival—is observed only at discharge. While clinicians can design intermediate rewards, their decision-making implicitly accounts for unmeasured variables and optimizes for latent objectives, including adverse event prevention, that are difficult to quantify. This motivates inverse reinforcement learning (IRL) to discover dense rewards from expert behavior. However, existing IRL methods are not designed for this setting: trajectory-level matching fails to capture situational clinical expertise, and standard approaches are affected by partial observability due to the lack of contextual modeling. We introduce **Context-IRL**, which extends Maximum Entropy IRL through three key designs: (1) a U-Net architecture that infers dense contextual rewards by integrating information across long trajectories, (2) a sampled-Q-matching method that leverages hindsight and aggregates forward-looking returns along trajectories, and (3) transition-level maximum entropy modeling with action infilling that provides stable learning signals. We evaluate Context-IRL on dual vasopressor control for septic shock across three real-world datasets. Experiments demonstrate that auxiliary rewards discovered by Context-IRL substantially improve RL training, yielding over 40% improvement over IRL baselines under off-policy evaluation.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

In medical applications, offline reinforcement learning (RL) has been widely applied to decision support for life-threatening conditions such as septic shock (Kalimouttou et al., 2025; Saria, 2018; Komorowski et al., 2018; Choi et al., 2024; Jayaraman et al., 2024; Yu & Huang, 2023; Adams et al., 2022). These methods learn from treatment trajectories where clinicians administer vasopressors (e.g., norepinephrine, vasopressin) to maintain hemodynamic stability over extended ICU stays. The sequential complexity of these decisions has motivated RL-based approaches that recommend actions to assist critical-care decision-making (Hunt et al., 1998; Garg et al., 2005).

Nevertheless, a major barrier to learning effective medical RL policies is reward sparsity: clinical outcomes are observed at discharge, exacerbating temporal credit assignment challenges (Minsky, 1961; Sutton, 1984; 1988; Sutton et al., 1998). In practice, clinicians are often required to define dense auxiliary rewards for the RL systems designed to assist them (Peng et al., 2018; Raghu et al., 2017; Kalimouttou et al., 2025; Choi et al., 2024), bringing non-trivial challenges to domain experts whose expertise lies in patient care.

To tackle this issue, we turn to inverse reinforcement learning (IRL) (Ng et al., 2000; Abbeel & Ng, 2004; Ziebart et al., 2008) to discover rewards from observed trajectories. One line of work in IRL builds maximum entropy formulations with trajectory-level reward feature-matching (Ziebart et al., 2008). It is later extended with Guided Cost Learning which uses a sampling approach (Finn et al., 2016). (Garg et al., 2021) adapted the approach to model-free Q-learning. However, trajectory-level maximum entropy IRL methods could struggle with the full complexity of long sequences. An alternative is transition-level autoregressive modeling, which captures the conditional action distribution at each timestep (Schroecker et al., 2019; Papamakarios et al., 2017). By operating at the transition level, these methods can capture finer-grained dynamics, making them well-suited to uncover latent factors and rewards in sparse-outcome environments. Another line of work follows the generative

adversarial approach, such as GAIL (Ho & Ermon, 2016). AIRL (Fu et al., 2017) demonstrates that rewards can also be recovered, though adversarial methods can be unstable over long horizons. Finally, related work has explored reward shaping and auxiliary reward learning to improve RL training (Suay et al., 2016; Brys et al., 2015). We adopt this auxiliary reward perspective for applying RL in septic shock treatment.

Furthermore, applying existing IRL methods to clinical decision-making presents additional challenges. First, clinical states are inherently partially observable: aspects of patient state that influence decisions, such as visual appearance, are not systematically recorded. Traditional IRL methods that infer rewards solely from variables observed at the current timestep struggle in this setting, as they fail to account for unobserved factors and information from other timesteps. Second, clinical trajectories span multiple days with hundreds to thousands of timesteps. This long-horizon setting poses challenges for trajectory-level methods that model entire episodes and exacerbates training instabilities in adversarial approaches.

To address these challenges, we introduce **Context-IRL** (*Contextual Inverse Reinforcement Learning*), a method designed for partial observability and long horizons in clinical trajectories. Our first insight is that reward inference improves by leveraging both backward and forward-looking observations to reveal hidden aspects of patient state—hence we term the learned rewards contextual. Our second insight is that reward discovery benefits from modeling per-timestep action distributions conditioned on trajectory context. Given this conditional distribution, latent reward functions can be shaped by predicting transition-level actions, exploiting the fact that action selection is weighted by the Q-function—aggregated future rewards.

Based on these insights, we introduce the Context-IRL and RL two-stage system (Figure 1). In the IRL step, a U-Net architecture processes multi-day trajectories and generates per-timestep reward values. These rewards are aggregated into sampled Q-values representing discounted future returns. Since action selection is weighted by Q-values, we learn the reward function by optimizing a loss that infills expert actions. In the RL step, the discovered rewards serve as auxiliary signals alongside sparse clinical outcomes. During policy optimization, the learned policy can refine beyond expert actions. We evaluate this system on vasopressor control for septic shock in the ICU, where clinicians maintain hemodynamic stability over multi-day trajectories to ensure patient survival.

To summarize, the main contributions of this paper are:

- **Contextual reward inference with a U-Net for multi-day trajectory modeling**: We adopt a U-Net architecture with forward and backward-looking capabilities for contextual reward inference. The network processes ICU trajectories spanning multiple days with hundreds of timesteps, enabling long-range contextual information propagation to generate dense rewards and facilitate credit assignment in sparse-outcome environments.

- **IRL through action trajectory infilling with sampled-Q-matching**: We extend Maximum Entropy IRL by defining the action probability at each timestep as a function of sampled Q-values computed from inferred contextual rewards. This formulation yields a stable prediction loss for learning a reward model through action trajectory infilling.

- **Empirical validation on septic shock treatment**: We evaluate Context-IRL on vasopressor control across three clinical datasets (eICU, MIMIC-IV, UPMC), achieving 46% improvement in expected reward over existing methods.

2. Related Work

Reinforcement learning has been applied extensively to septic shock treatment. (Saria, 2018) proposed individualized sepsis treatment using RL, while (Komorowski et al., 2018) provided an early comprehensive study using publicly available datasets. More recently, (Choi et al., 2024) offered in-depth analysis and stable training procedures for deep RL in sepsis, and (Jayaraman et al., 2024) reviewed RL applications in ICU settings. (Kalimouttou et al., 2025) presented a focused study of vasopressin initiation for patients on first-line vasopressors, with comprehensive cross-institution evaluation.

For related Q-learning and RL literature, we briefly review deep Q-learning (DQN)(Mnih et al., 2015) and offline RL methods such as conservative Q-learning(Kumar et al., 2020) and implicit Q-learning (Kostrikov et al., 2021). Model-based approaches have also been applied to offline RL (Kidambi et al., 2020; Yu et al., 2020; 2021) by learning transition dynamics.

Inverse reinforcement learning was introduced in the Maximum Entropy IRL framework (Ziebart et al., 2008), which provided a principled probabilistic approach to resolve ambiguity in reward inference. Boularias et al. (2011) extended this framework with a model-free algorithm based on minimizing relative entropy between trajectory distributions. Deep IRL (Wulfmeier et al., 2015) subsequently applied neural networks to approximate nonlinear reward functions within the maximum entropy paradigm. Das et al. (2020)

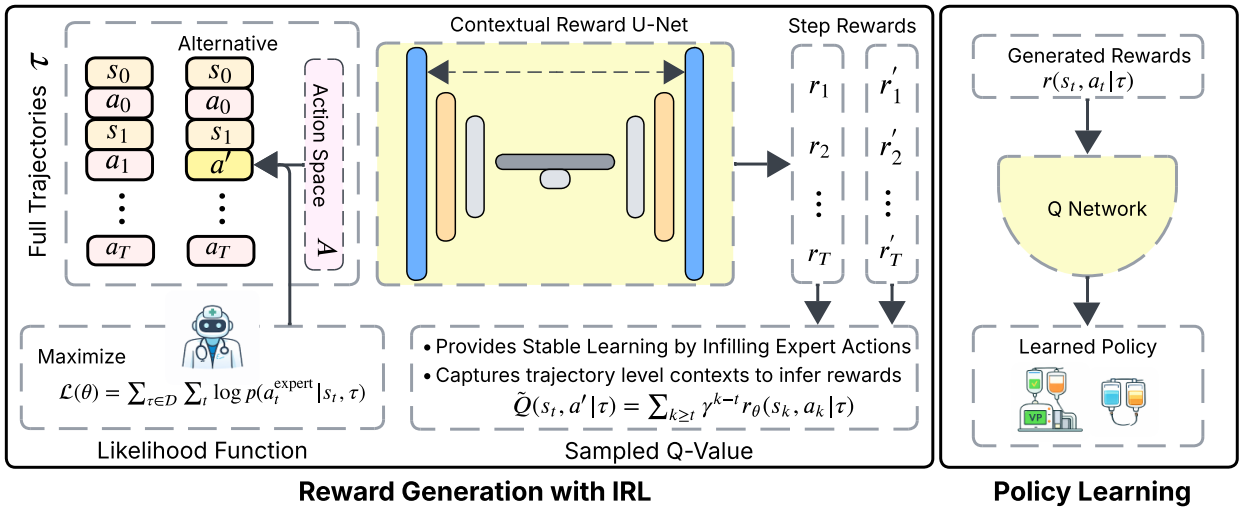


Figure 1. Context-IRL and RL Two Stage System with Model Architecture

further extended these methods to model-based RL with visual control tasks. Guided Cost Learning (Finn et al., 2016) formulated joint optimization of policy and reward using sample-based approximations for MaxEnt inverse optimal control. IQ-Learn (Garg et al., 2021) developed an IRL method compatible with Q-learning in both offline and online settings, avoiding adversarial training. These methods are closely related to imitation learning; notably, Generative Adversarial Imitation Learning (GAIL) (Ho & Ermon, 2016) connected IRL with generative adversarial networks for model-free imitation in high-dimensional environments. Arora & Doshi (2021) provided an overview of IRL methods, challenges, and applications.

3. Problem Definition

We consider the problem of learning policies given a Markov Decision Process (MDP) given historical data $\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\}$. Here the MDP is denoted $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition dynamics, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function.

In the case of learning the optimal policy for vasopressor administration in sepsis, the state $s_t \in \mathcal{S} \subseteq \mathbb{R}^d$ comprises patient information, physiological measurements[1], and organ support[2]. The action space $\mathcal{A} = \{(vp_1, vp_2)\}$ describes dual vasopressor administration, where $vp_1 \in \{0, 1\}$ represents the use of vasopressin (binary) and $vp_2 \in (0, 0.5]$ represents norepinephrine dose (mcg/kg/min). To align with clinical practice, we discretize vp_2 into $N \in \{5, 10\}$ evenly-spaced bins with values at bin centers: $vp_2 \in \{(n + \frac{1}{2})\delta \mid n = 0, 1, \ldots, N - 1\}$, where δ is the bin width.

For clinician-designed rewards (**CD rewards**), we use the reward function defined in Kalimouttou et al. (2025), which captures both patient mortality and intermediate indicators of clinical progress. As shown in Table 1, the reward is dominated by a mortality penalty, supplemented by bonuses for survival at each timestep, improved clinical parameters (mean blood pressure, lactate levels, and Sequential Organ Failure Assessment scores), and reduced norepinephrine usage.[3] Notably, this reward function, like many in healthcare, is essentially sparse, as the mortality signal occurs only at the end of the patient trajectory.

Component	Reward
In-hospital mortality (terminal)	-20
Base (each transition)	$+1$
MBP rising to $\geq$65 mmHg (next 4h window)	$+1$
Lactate decrease (next 6h window)	$+1$
SOFA score decrease (next 6h window)	$+3$
Norepinephrine reduction (next 4h window)	$+1$

Table 1. Reward Function Components.

4. Preliminaries

We formulate the Q-learning for offline RL, with temporal difference (TD) learning. Central to this method is the

[1]The measurements include mean blood pressure (MBP), lactate levels, blood urea nitrogen, serum creatinine, (total) fluid, urine of the hour, corticosteroid, and Sequential Organ Failure Assessment (SOFA) score.

[2]The support includes mechanical ventilation and renal replacement therapy.

[3]See Appendix D for a detailed reward description.

Q-function $Q^\pi(s,a)$, which represents the expected cumulative reward by taking action a in state s and following policy π thereafter:

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid s_t = s, a_t = a \right]. \quad (1)$$

The optimal Q-function satisfies the Bellman optimality equation:

$$Q^*(s,a) = \mathcal{R}(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} [\max_{a'} Q^*(s',a')].$$

Q-learning approximates Q^* by learning a function Q_θ with parameters θ that minimizes the temporal difference (TD) error:

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

where Q_{θ^-} is a target network updated periodically for stability. The learned policy is then defined as the argmax of the Q-function, i.e., $\pi^*(s) = \arg\max_a Q_\theta(s,a)$.

Learning from sparse rewards is challenging, motivating the use of inverse reinforcement learning to discover dense auxiliary signals. We formulate as context the Maximum Entropy IRL (Ziebart et al., 2008) algorithm, which assumes that experts optimally maximize reward while exhibiting maximum entropy behavior—that is, among all policies achieving the same expected reward, experts are least committal, making no assumptions beyond what is necessary to explain the demonstrations. Under this principle, the probability of an expert trajectory τ follows the distribution:

$$P(\tau) = \frac{\exp(\sum_\tau r)}{Z}, \quad \text{where } Z = \sum_{\tau \sim \mathcal{D}} \exp(\sum_\tau r). \quad (3)$$

The reward function is then learned by maximizing the likelihood of expert trajectories.

However, Maximum Entropy IRL has key limitations for clinical settings. First, it does not model how unobserved latent factors influence expert decisions at each timestep. To address this, we borrow from imputation: examining both past and future states to infer expert actions at a given timepoint. Second, the standard formulation depends directly on rewards, which are noisy in data-limited clinical settings. Rather than represent optimal policies in terms of rewards, we represent optimal actions in terms of the Q-function, aligning with statistically efficient Q-learning-based approaches. Third, trajectory probabilities in (3) are invariant to reward ordering—state-action pairs with identical rewards can be permuted without changing the trajectory probability, which may not reflect clinical practice where sequential order matters. For clinical settings like sepsis, we need a reward discovery method that captures transition-level dependencies to preserve temporal structure.

5. Context-IRL: Contextual Inverse Reinforcement Learning

We introduce Contextual Inverse Reinforcement Learning (Context-IRL) (Figure 1) to address three challenges common to clinical RL scenarios: (i) long trajectories with limited sample sizes and high variance, (ii) the need to capture transition-level dependencies and correlations, and (iii) partial observability due to unmeasured factors. We address these by extending Maximum Entropy IRL through two key design choices. First, we learn contextual rewards with a U-Net architecture that leverages forward and backward-looking observations to infer dense reward trajectories, mitigating partial observability. Second, we replace trajectory-level reward-feature matching with sampled-Q-matching to predict transition-level expert actions with hindsight, where Q-values are computed as observed forward-looking rewards along expert trajectories, providing a more stable learning signal for long-horizon clinical settings. The Context-IRL-learned rewards are then combined with sparse outcomes for policy learning. Additional details are provided in the Supplementary Materials.

5.1. Actional Trajectory Infilling with Sampled-Q-Matching

We apply the maximum entropy principle for the action distribution rather than the trajectory level. Rather than reward feature-matching (Ziebart et al., 2008), we employ Q feature-matching, which enables a representation of the action distribution using sampled Q-values.

First, we define the *sampled Q-value* as the discounted sum of future rewards along the observed trajectory:

$$\tilde{Q}(s_t, a_t|\tau) = \sum_{k \geq t} \gamma^{k-t} r_\theta(s_k, a_k|\tau), \quad (4)$$

which is the inner term in the expectation of (1). Following the principle of maximum entropy (Ziebart et al., 2008), we seek the distribution over actions at each timestep that is maximally non-committal while being consistent with the constraint that actions with higher Q-values should be preferred. Under this principle, the least committal distribution satisfying these constraints is the exponential family distribution. Consequently, at each timestep t, we model the expert's action distribution as:

$$p(a_t|s_t,\tau) \propto \exp(\tilde{Q}(s_t, a_t|\tau)), \quad (5)$$

where $\tau = (s_0, a_0, \ldots, s_T, a_T)$ denotes the full trajectory sequence providing context for reward computation.

We call this *sampled-Q-matching*, since it matches observed trajectory returns rather than expected Q-functions. Computing expected Q-functions would require taking expectations over all possible future trajectories under a learned policy, which is infeasible in offline settings without access to the state transition dynamics $P(s'|s,a)$ or the ability to simulate multiple rollouts from each state. This approach offers two advantages over reward-feature matching. First, by aggregating future rewards, $\tilde{Q}$ provides a more stable learning signal through variance reduction—critical for multi-day clinical trajectories with hundreds of timesteps. Second, sampled-Q-matching considers long-term consequences of actions rather than immediate rewards, which is better able to capture delayed effects as common in clinical settings. See Figure 3 for an illustration of the algorithm.

5.2. Contextual Reward Learning via U-Net

To evaluate Equation 4, we must learn the per-timestep reward function $r_\theta(s_t, a_t|\tau)$. A key challenge in clinical settings is partial observability: critical aspects of patient state may influence clinician decisions but are not systematically recorded. To address this, we leverage *context*—using the entire trajectory, including both past and future observations, to learn and infer dense rewards across all transitions.

We implement this through a U-Net architecture (Ho et al., 2020; Dhariwal & Nichol, 2021), originally developed for image generation, adapted here for long-horizon reward inference. The U-Net's key advantages are its skip connections that enable direct information flow across distant timesteps and its hierarchical encoder-decoder structure that captures multi-scale temporal patterns. Here we use the U-Net to process the complete trajectory, generating per-timestep rewards $r_\theta(s_t, a_t|\tau)$ conditioned on the entire state/action sequence $\tau = (s_0, a_0, \ldots, s_T, a_T)$ to better infer the hidden patient state and decision quality at time t. To evaluate action candidate a' at timestep t, we construct a modified trajectory by replacing the expert action a_t with a': $(s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t, a', s_{t+1}, a_{t+1}, \ldots, s_T, a_T)$. The U-Net processes this modified trajectory through an encoder that condenses the sequence into an embedding space, then through a decoder that propagates information back to the full horizon, producing rewards for all timesteps. The hierarchical encoder-decoder structure with skip connections enables information propagation across distant timesteps—critical for multi-day ICU trajectories where early vasopressor decisions influence outcomes days later. Given rewards $\{r_\theta(s_k, a_k|\tau)\}_{k=t}^{T}$, we compute the sampled Q-value as $\tilde{Q}(s_t, a'|\tau) = \sum_{k \geq t} \gamma^{k-t} r_\theta(s_k, a_k|\tau)$.

The training objective maximizes the log-likelihood of expert actions across all timesteps via backpropagation through the U-Net:

$$\mathcal{L}(\theta) = \sum_{\tau \in \mathcal{D}} \sum_{t} \log p(a_t^{\text{expert}}|s_t, \tau),$$

where $\mathcal{D}$ denotes the dataset of expert trajectories collected from retrospective clinical data. We train the U-Net using a replay buffer. However, unlike standard RL replay buffers that store individual transitions, Context-IRL requires sequential context so a sequential replay buffer is used instead. That is, we maintain a buffer $\mathcal{RB} = \{\mathcal{S}_e\}$ that stores complete treatment episodes $\mathcal{S}_e = \{(s_t, a_t, r_t, s_{t+1})_{t \in e}\}$ and sample trajectories ending at least D timesteps before episode termination so that there is sufficient future context for computing sampled Q-values.

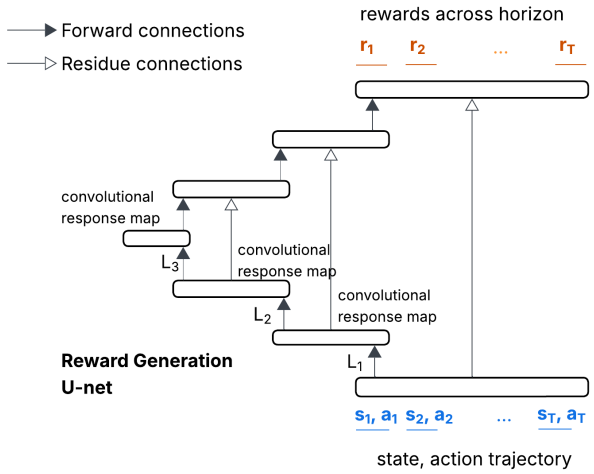


Figure 2. Contextual U-Net Architecture for Reward Generation. The U-Net processes full trajectories to generate per-timestep rewards $r_\theta(s_t, a_t)$ that incorporate both past and future context, enabling hindsight reward imputation under partial observability.

5.3. Combining IRL with Sparse Rewards

As common in healthcare, the reward function is dominated by a few critical, sparse signals such as patient mortality. It is thus beneficial to include these critical signals in the IRL procedure to further improve credit assignment across long trajectories. To this end, we introduce a semi-supervised variant, **SS-C-IRL**, that augments the contextual reward U-Net with learned temporal propagation of the mortality signal. Specifically, we apply a trainable transposed convolutional neural network that takes the scalar mortality outcome (1 for survival, 0 for death) and propagates it across the trajectory length, learning a position-dependent weight function for temporal credit assignment.

Formally, let $m \in \{0, 1\}$ denote the mortality outcome and, let $f_\phi : \mathbb{R} \times \mathbb{N} \to \mathbb{R}^T$ denote the transposed convolutional

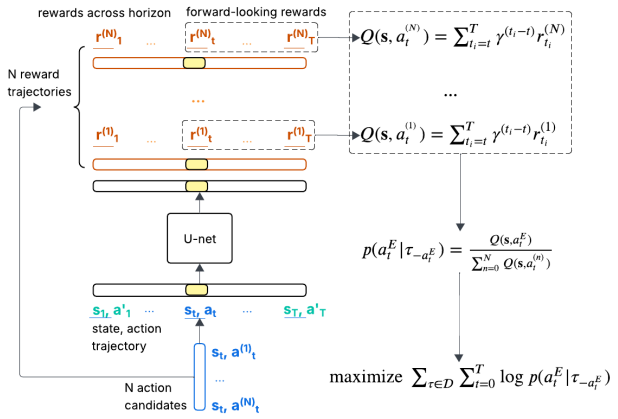


Figure 3. Illustration of sample-Q-matching. Given N action candates at time t, the U-Net infers N reward trajectories across the horizon. The forward-looking rewards are aggregated to be the sample-Q-value. The Q-values across N action candidates are weighed to represent the an exponential family distribution over action a_t. The infill prediction objective can then be formed to learn the reward function.

network parameterized by ϕ, which maps the scalar mortality signal and trajectory length T to a sequence of temporal weights. The final reward at each timestep combines the IRL-learned rewards with the propagated mortality signal:

$$r_{\text{total}}(s_t, a_t|\tau) = r_\theta(s_t, a_t|\tau) + f_\phi(m, T)_t.$$

The transposed convolutional network learns how to temporally distribute the mortality signal during training. Unlike uniform propagation, this allows the model to learn that decisions at certain trajectory positions (e.g., early critical interventions vs. late weaning decisions) may have different relationships to the final outcome. The parameters ϕ are jointly optimized with the U-Net parameters θ via the transition-level maximum entropy objective (Equation 5).

This semi-supervised approach provides two benefits: (1) it explicitly incorporates the known terminal outcome into credit assignment, and (2) it allows the model to learn trajectory-phase-dependent reward shaping from the mortality signal.

5.4. Downstream Q-Learning for Policy Deployment

The rewards learned by Context-IRL depend on full trajectory context: $r_\theta(s_t, a_t|\tau)$. While this enables hindsight reward imputation during training on retrospective data, it poses a challenge for deployment: when treating a new patient, future observations are not yet available. To obtain a deployable policy, we use the discovered rewards to train a standard Q-function $Q_\psi(s, a)$ via offline Q-learning.

Specifically, after Context-IRL learns the reward function r_θ, we apply it to all expert trajectories to generate reward labels. We then train a Q-network $Q_\psi(s, a)$ using temporal difference (TD) learning as in Equation 2.

The learned Q-function Q_ψ depends only on current state and action, enabling deployment via the greedy policy $\pi(s) = \arg\max_a Q_\psi(s, a)$. This downstream Q-learning step also enables fair comparison with baselines: both Context-IRL and baseline methods use the same Q-learning framework, isolating the effect of reward discovery.

6. Experiments

Dataset and Preparation. We study three ICU datasets containing adult patients who met Sepsis-3 shock criteria: *UPMC*: 6,251 patients from 18 hospitals in Pennsylvania; *MIMIC-IV* (Johnson et al., 2023): 3,056 patients from an academic medical center in Boston; *eICU-CRD* (Pollard et al., 2018): 910 patients from 208 hospitals in the U.S.

Irregular Electronic Health Record (EHR) data were transformed into trajectories as follows. *Temporal discretization*: patient stays were discretized into 1-hour transitions; trajectories began after shock onset and were truncated at 120 hours, ICU discharge, or shock recovery. *Features*: each transition captured 14 time-varying clinical features including vital signs, lab results, and organ support status. *Missing data*: we employed Last-Observation-Carried-Forward (LOCF) imputation.

Evaluation Methodology. We adopt off-policy evaluation (OPE) with weighted importance sampling (WIS) on the held-out test set as the primary evaluation method. The method requires extraction of model-recommended actions. For the Q-learning baseline, actions are extracted from the model trained directly on true rewards. For IRL methods including Context-IRL, WIS OPE is performed on actions recommended by the downstream Q-learning model trained on IRL-discovered rewards. To evaluate with WIS OPE, we first identify optimal actions recommended by an RL model and train a softmax classifier to represent the target policy (π_m) and train a softmax classier to represent the baseline policy (π_c). Then, the WIS evaluations are performed with Equation 6.

$$R_{\text{WIS}} = \frac{\sum_{j=0}^{N} w_j R}{\sum_{j=0}^{N} w_j}, \quad w_j = \prod_{t_i=0}^{j} \frac{\pi_m(a_{t_i}^{\text{cli.}} \mid s_{t_i})}{\pi_c(a_{t_i}^{\text{cli.}} \mid s_{t_i})}, \tag{6}$$

6.1. Q-Learning Baseline

We train a Q-learning model directly on the reward function defined in Table 1 to establish a baseline for IRL comparison. The model is trained on sepsis treatment data from eICU (Pollard et al., 2018) and MIMIC-IV (Johnson et al.,

2023)[4] for 500 epochs and evaluated with WIS OPE.

For comparison, we also include the *Dual Mixed* setting where vp_1 is a binary variable indicating vasopressin initiation (Kalimouttou et al., 2025) and vp_2 is a continuous variable. Table 2 shows the evaluation results. Finer discretization in the dual discrete (Dual Disc.) action space yields better performance. Both discrete action space models show over $3\times$ improved WIS expected reward compared with the Dual Mixed model.

Table 2. **Importance Sampling (IS) and Weighted Importance Sampling (WIS) OPE Results.** Baseline (Base.) is the average reward per trajectory across the test-set. Trajectory-level (Traj. Lvl) WIS is shown, along with the improvement in trajectory-level WIS (Imp.) between model-recommended actions and clinician actions. The 95% Confidence Interval (CI) for WIS improvement is shown in parentheses, obtained by bootstrapping across test set trajectories.

Model	Config	Traj. Lvl		Δ	95%CI
		Base.	WIS		
Binary vp_1	–	112.01	114.6	2.6	[-8.2, 12.2]
Dual Mixed	–	–	118.9	6.8	[-0.8, 14.2]
Dual Disc.	3 bins	–	122.3	10.3	[-1.1, 23.7]
Dual Disc.	5 bins	–	137.8	25.7	[9.8, 41.2]
Dual Disc.	10 bins	–	137.5	25.5	[12.9, 38.7]

6.2. Comparison of IRL Methods

We evaluate Context-IRL and SS-C-IRL against established IRL baselines on the eICU and MIMIC-IV datasets. All methods use the same training, validation, and test splits to ensure fair comparison. Baseline IRL methods—Maximum Entropy IRL (Ziebart et al., 2008), Guided Cost Learning (Finn et al., 2016), and IQ-Learn (Garg et al., 2021)—are implemented with identical reward network architectures. Maximum Entropy IRL applies a per-transition cost function, GCL uses the IOC version that jointly learns policy and cost functions, and IQ-Learn is evaluated with chi-squared, KL divergence, and Jensen distance metrics. Hyperparameters for each baseline are selected based on validation set performance.

After each IRL algorithm recovers its reward function $R(s, a)$, we apply the same downstream Q-learning procedure (Section 4.4) with identical Q-network architecture and hyperparameters across all methods. This ensures that performance differences reflect the quality of discovered rewards rather than differences in downstream RL training. For Context-IRL, we use a U-Net architecture with a 3-layer convolutional encoder, 3-layer transposed convolutional decoder, and residual connections between corresponding encoder-decoder layers. Each layer has 64 hidden units. SS-C-IRL augments this with the mortality propagation network described in Section 4.3.

All methods are evaluated using Weighted Importance Sampling (WIS) off-policy evaluation on the held-out test set, with the clinician trajectory baseline average reward $E[R] = 112.01$ (all methods use the same test set, hence identical baseline values). Evaluation uses a discrete action space with 5 vp_2 bins. Table 3 presents the comparison.

Results are evaluated with a discrete action space using 5 vp_2 bins. Table 3 presents the comprehensive comparison.

Table 3. Comparison of IRL Methods Using WIS Off-Policy Evaluation on eICU and MIMIC-IV Datasets. All methods use downstream Q-learning with identical architectures and hyperparameters. Standard errors computed via bootstrapping with 1,000 resamples. Higher Traj. WIS and Imp. values indicate better performance.

Method	Base.	Traj. WIS	Δ	95%CI
Maximum Entropy IRL	112.01	119.1	7.0	[-3.5, 18.8]
Guided Cost Learning	–	117.4	5.4	[7.5, 35.1]
IQ-Learn	–	132.7	20.8	[7.5, 35.1]
Context-IRL (this work)	–	**140.5**	**28.4**	[13.3, 45.0]
SS-C-IRL (this work)	–	138.1	26.1	[8.7, 44.7]

Context-IRL substantially outperforms all baseline IRL methods, achieving $1.25\times$ improvement over the best baseline. Notably, Context-IRL learns rewards purely from expert demonstrations without access to the mortality signal, yet outperforms Q-learning trained on rewards that explicitly include mortality. This demonstrates that contextual trajectory modeling successfully captures implicit expert decision-making logic. SS-C-IRL achieves $1.23\times$ improvement over the baseline by incorporating the sparse terminal mortality signal. The trajectory-level methods (MaxEnt IRL and GCL) generally perform worse than transition-level methods, likely due to challenges in credit assignment over long clinical trajectories. IQ-Learn, a transition-level method, performs better than trajectory-level approaches but still falls short of Context-IRL, which benefits from contextual reward inference and hindsight action infilling.

6.3. Evaluating Generalizability

We now assess the generalizability of the learned model. The models are trained on the eICU and MIMIC-IV datasets, and the UPMC dataset is completely held-out, used for validation and test sets. The validation and test sets are of equal sizes. The results in Table 4 show that using Context-IRL substantially improves over existing methods, demonstrating a 22% improvement over the RL model trained directly with clinician-designed rewards. Context-IRL also improves 25% over Max-Ent IRL. The SS-C-IRL method shows the largest improvements. It demonstrates a 43% improvement

[4] See Appendix A for data and processing details.

over RL model trained on CD rewards, and a **46%** improvement over Max-Ent IRL, the best performing IRL baseline. These results suggest that accounting for latent factors that clinician behavior may be a critical component for improving out-of-sample generalizability of RL methods.

Table 4. Combined Evaluation of IRL Methods on eICU, MIMIC-IV, and UPMC Datasets. The table shows Clinician (Cli.) and Model (Mod.) actions evaluated with WIS OPE. The table evaluates the trajectory-level rewards using WIS; Blue highlights model learned with the original rewards, orange highlights model learned with Context-IRL learned rewards without observing mortality.

WIS: CD Reward	WIS		Δ	95% CI
IRL Method	Cli.	Mod.		
CD Rewards	92.7	150.5	57.8	[34.0, 79.8]
Max-Ent IRL	-	149.2	56.5	[40.3, 72.6]
GCL	-	141.7	49.0	[36.0, 62.0]
IQ-Learn	-	145.7	53.0	[28.4, 72.8]
Context-IRL	-	163.4	70.7	[44.0, 95.4]
SS-C-IRL	-	**175.4**	**82.7**	[60.2, 101.9]

6.4. Interpreting Learned Rewards

We compute Spearman rank correlation between IRL-discovered rewards and the survival reward from clinician-designed rewards. Trajectory-level rewards are computed using each reward function, summed, and normalized by trajectory length. The values are then ranked to compute Spearman correlation. Results are shown in Table 5 along with Pearson correlation. Context-IRL and SS-C-IRL exhibit strong per-trajectory correlations with the survival reward.

Table 5. Spearman Rank Correlation and Pearson Correlation (IRL Reward vs. the Survival (-Mortality) Reward).

Method	Spearman	Pearson
Max-Ent	-0.066	0.002
GCL	-0.100	-0.044
IQ-Learn	-0.121	-0.409
Context-IRL	0.198	0.299
SS-C-IRL	**0.194**	**0.347**

Next, Canonical Correlation Analysis (CCA) is used to identify canonical pairs correlating IRL-learned rewards with the 5 components of clinician-defined rewards.[5] CCA is performed on the union of the validation and test sets, expressed as $X^* = \sum_{i=1}^{5} \lambda_i Y_i$, where X is the IRL reward score for a trajectory and Y_i is the ith component of the original reward function. Both X and Y are normalized by trajectory length.

From the CCA analysis, we can see clear correlation between the IRL reward and the survival (-mortality) reward. some per-component coefficients are neutral (e.g. Lactate, BP, etc) for Context-IRL learned rewards. The rule from clinician design may not be exactly matching of the rewards discovered by Context-IRL, while eventually the reward signals conditioned the Q-learning model to obtain statistically significant WIS OPE improvements.

[5]Including base reward of +1 for each timestep.

Table 6. CCA Component Coefficients (λ): $X^* = \sum_{i=1}^{5} \lambda_i Y_i$

	Component (λ_i)					
Method	ρ	Lact.	BP	SOFA	Norepi.	Surv.
MaxEnt IRL	0.077	−0.01	−0.04	−0.57	0.77	−0.26
GCL	0.236	0.33	0.00	−0.80	−0.45	−0.20
IQ-Learn	0.503	−0.10	0.10	0.01	0.67	−0.73
Context-IRL	0.493	−0.17	−0.10	0.01	−0.64	**0.74**
SS Context-IRL	**0.530**	−0.14	−0.10	−0.03	−0.61	**0.78**

Note: Base = Survival, Lact. = Lactate, BP = Blood Pressure, Norepi. = Norepinephrine, Mort. = Mortality.

7. Conclusion

We propose Contextual Inverse Reinforcement Learning (Context-IRL) to discover dense auxiliary rewards for RL in septic shock management, addressing the fundamental challenges of reward sparsity and partial observability in clinical trajectories. Context-IRL leverages a U-Net architecture with bidirectional information propagation to process multi-day ICU trajectories spanning hundreds of timesteps. This enables contextual reward inference that accounts for both forward and backward-looking observations, capturing hidden factors that influence decisions. We introduce a sampled-Q-matching method that defines action probabilities as a function of aggregated future rewards, yielding a stable prediction loss for learning reward functions through action trajectory infilling. This formulation improves upon trajectory-level methods and addresses the training instabilities of adversarial approaches.

Combined with a discrete action space, we develop a two-stage Context-IRL to RL system for dual vasopressor control in septic shock treatment. The discovered rewards serve as auxiliary signals alongside sparse clinical outcomes, enabling policy optimization that can refine beyond observed expert actions. We validate Context-IRL across three real-world clinical datasets (eICU, MIMIC-IV, UPMC), demonstrating robust generalization across institutions. The combined system achieves over 46% improvement in expected reward under weighted importance sampling off-policy evaluation compared to existing IRL baselines. These contributions advance a blueprint for clinical decision support systems that effectively recover dense auxiliary rewards in sparse-outcome environments and integrate seamlessly into critical medical workflows.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Adams, R., Henry, K. E., Sridharan, A., Soleimani, H., Zhan, A., Rawat, N., Johnson, L., Hager, D. N., Cosgrove, S. E., Markowski, A., et al. Prospective, multi-site study of patient outcomes after implementation of the trews machine learning-based early warning system for sepsis. *Nature medicine*, 28(7):1455–1460, 2022.

Arora, S. and Doshi, P. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021. arXiv preprint arXiv:1806.06877.

Boularias, A., Kober, J., and Peters, J. Relative entropy inverse reinforcement learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15 of *Proceedings of Machine Learning Research*, pp. 182–189, Fort Lauderdale, FL, USA, 2011. PMLR.

Brys, T., Harutyunyan, A., Suay, H. B., Chernova, S., Taylor, M. E., and Nowé, A. Reinforcement learning from demonstration through shaping. In *IJCAI*, pp. 3352–3358, 2015.

Choi, Y., Oh, S., Huh, J. W., Joo, H.-T., Lee, H., You, W., Bae, C.-m., Choi, J.-H., and Kim, K.-J. Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records. *Communications Medicine*, 4(1): 245, 2024.

Das, N., Bechtle, S., Davchev, T., Jayaraman, D., Rai, A., and Meier, F. Model-based inverse reinforcement learning from visual demonstrations. In *Proceedings of the 4th Conference on Robot Learning (CoRL)*, volume 155 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Finn, C., Levine, S., and Abbeel, P. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *JMLR: W&CP*, pp. 49–58, New York, NY, USA, 2016.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., Sam, J., and Haynes, R. B. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Jama*, 293(10): 1223–1238, 2005.

Garg, D., Chakraborty, S., Cundy, C., Song, J., and Ermon, S. Iq-learn: Inverse soft-q learning for imitation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pp. 4565–4573, 2016.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hunt, D. L., Haynes, R. B., Hanna, S. E., and Smith, K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *Jama*, 280(15):1339–1346, 1998.

Jayaraman, P., Desman, J., Sabounchi, M., Nadkarni, G. N., and Sakhuja, A. A primer on reinforcement learning in medicine for clinicians. *NPJ Digital Medicine*, 7(1):337, 2024.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Kalimouttou, A., Kennedy, J. N., Feng, J., Singh, H., Saria, S., Angus, D. C., Seymour, C. W., and Pirracchio, R. Optimal vasopressin initiation in septic shock: the oviss reinforcement learning study. *Jama*, 333(19):1688–1698, 2025.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33: 21810–21823, 2020.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33: 1179–1191, 2020.

Minsky, M. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961. doi: 10.1109/JRPROC.1961.287775. URL https://ieeexplore.ieee.org/document/4066245.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Ng, A. Y., Russell, S., et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Papamakarios, G., Pavlakou, T., and Murray, I. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

Peng, X., Ding, Y., Wihl, D., Gottesman, O., Komorowski, M., Lehman, L.-w. H., Ross, A., Faisal, A., and Doshi-Velez, F. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*, volume 2018, pp. 887, 2018.

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., and Ghassemi, M. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.

Saria, S. Individualized sepsis treatment using reinforcement learning. *Nature medicine*, 24(11):1641–1642, 2018.

Schroecker, Y., Vecerik, M., and Scholz, J. Generative predecessor models for sample-efficient imitation learning. *arXiv preprint arXiv:1904.01139*, 2019.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J.-L., and Angus, D. C. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, 02 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.0287. URL https://doi.org/10.1001/jama.2016.0287.

Suay, H. B., Brys, T., Taylor, M. E., and Chernova, S. Learning from demonstration for shaping through inverse reinforcement learning. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pp. 429–437, 2016.

Sutton, R. S. *Temporal Credit Assignment in Reinforcement Learning*. PhD dissertation, University of Massachusetts, Amherst, Amherst, MA, 1984. URL https://scholarworks.umass.edu/dissertations/AAI8410337/. Available from ProQuest Dissertations, AAI8410337.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1): 9–44, August 1988. doi: 10.1007/BF00115009. URL http://incompleteideas.net/papers/sutton-88-with-erratum.pdf.

Sutton, R. S., Barto, A. G., et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Wulfmeier, M., Ondruska, P., and Posner, I. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

Yu, C. and Huang, Q. Towards more efficient and robust evaluation of sepsis treatment with deep reinforcement learning. *BMC medical informatics and decision making*, 23(1):43, 2023.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

Yu, T., Kumar, A., Rafailov, R., Rajeswaran, A., Levine, S., and Finn, C. Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34:28954–28967, 2021.

Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 1433–1438. AAAI Press, 2008.

Appendix

A. Data Source and Processing

The patient features extracted from the eICU and MIMIC-IV databases serve as the state representation for the reinforcement learning (RL) algorithm, and the action space consists of vasopressin administration (binary) and norepinephrine dosing in the continuous range $(0, 0.5]$ (mcg/kg/min). The data are preprocessed such that missing values are imputed using forward filling, and norepinephrine dosages are clipped to the valid clinical range $(0, 0.5]$. We include unique patients experiencing their first episode of septic shock, defined using the Sepsis-3 criteria (Singer et al., 2016) and who were already receiving norepinephrine at the moment shock onset was identified. Shock onset could occur either while the patient was still in the emergency department or after ICU admission.

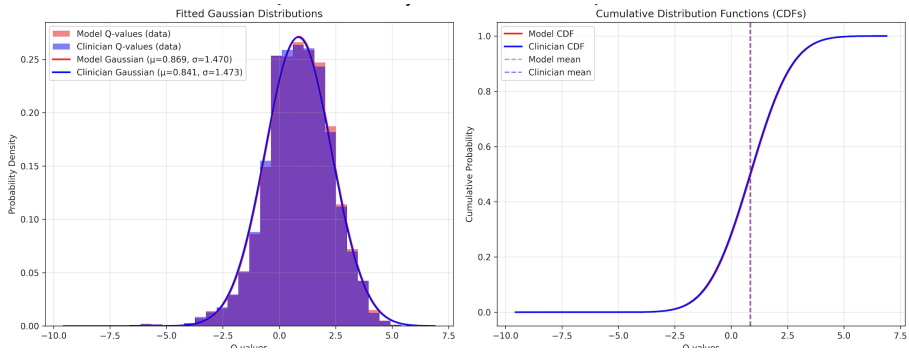
B. Fitted Q-Evaluation (FQE)

To evaluate the learned policies against historical clinician decisions without online deployment, we employ Fitted Q-Evaluation (FQE) with Gaussian distribution modeling. This approach computes Q-values for both the learned policy and clinician actions on the held-out test set, enabling direct comparison of expected returns. For each trained model, we extract Q-values by evaluating state-action pairs from the test trajectories where actions are either from the learned policy (model Q-values) or from historical clinician decisions (clinician Q-values). The resulting Q-value distributions are then fitted with Gaussian distributions using maximum likelihood estimation, providing parametric representations characterized by mean (μ) and standard deviation (σ) for both the model and clinician policies. This parametric approach enables computation of key metrics including the mean improvement gap (difference in expected returns), probability of improvement (likelihood that model Q-values exceed clinician Q-values), and effect size (Cohen's d) to quantify the magnitude of policy differences.

The FQE analysis also generates visualizations including overlapping histograms with fitted Gaussian curves, cumulative distribution functions (CDFs) showing the probability mass distribution of Q-values, and improvement probability curves that illustrate the likelihood of the model outperforming various Q-value thresholds. The probability of improvement at the clinician's mean Q-value serves as a critical metric, indicating how often the learned policy is expected to achieve better outcomes than historical treatment decisions. Additionally, the analysis computes Cohen's d as a standardized effect size measure. The larger the value of Cohen's d, the more meaningful are the difference between the means of two groups. This framework provides quantitative evidence for policy improvement while accounting for the inherent uncertainty in Q-value estimates. The FQE analysis was performed for policy performance across different model architectures and hyperparameter configurations. The histogram and sigmoid CDF visualizations are shown in Appendix C.

C. Fitted Q-Evaluation Illustrations

Figure 4. FQE comparison between learned patient model and clinician baseline (Binary vp_1).

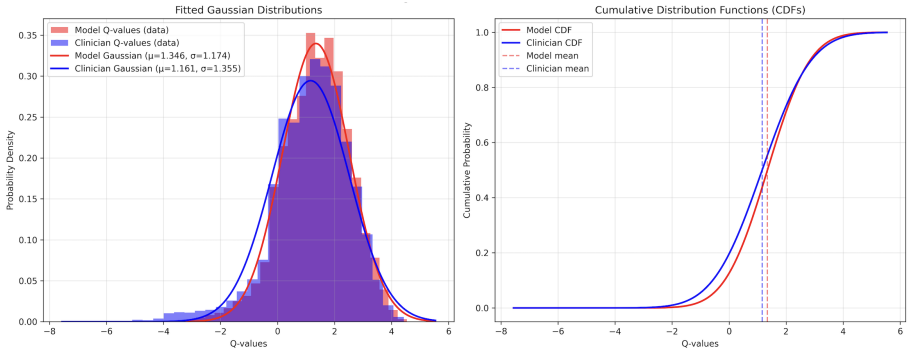


Binary vp_1 Model: distribution of Q-values from model (red) and clinician (blue).

D. Comprehensive Reward Function

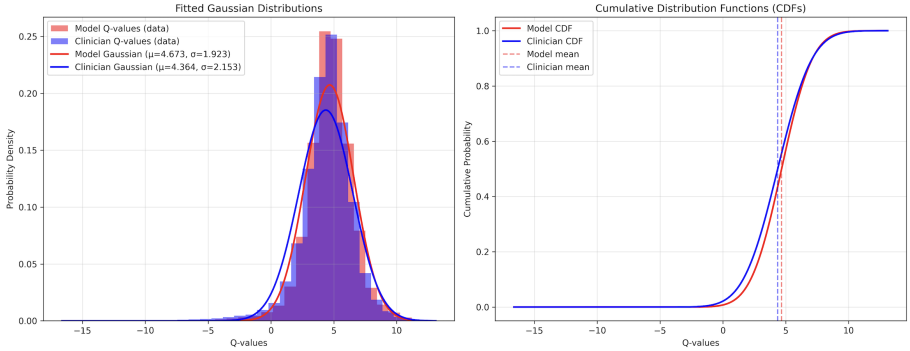
We directly follow the OVISS (Kalimouttou et al., 2025) study to implement a full and comprehensive reward function with adjusted rewards. The reward measures patient progress and response to vasopressor administration in a quantifiable

Figure 5. FQE comparison between the learned patient model and clinician baseline (Dual Mixed).



Dual Mixed Model: distribution of Q-values from model (red) and clinician (blue).

Figure 6. FQE comparison between the learned patient model and clinician baseline (Dual BD).



Block Discrete Model (10 bins): distribution of Q-values from model (red) and clinician (blue).

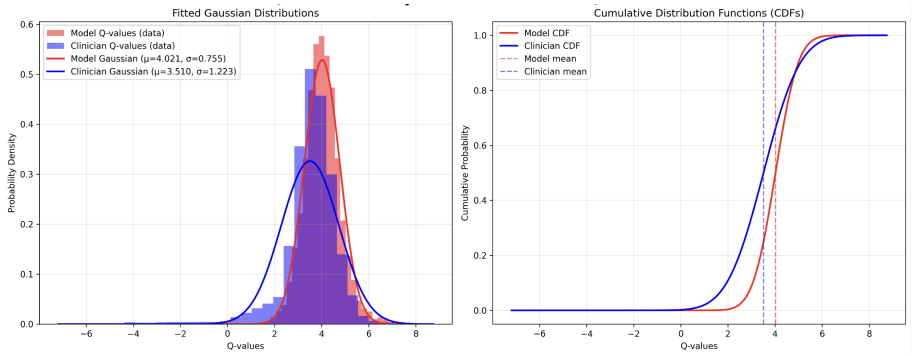
manner. The reward definition involves calculating an "adjusted reward" for each patient at different time points, which is a composite measure reflecting both survival and physiological benefits.

- **Survival Reward.** Each patient receives a base reward of $+1$ at every time step, reflecting the fundamental importance of maintaining survival.
- **Death Penalty.** If in-hospital mortality occurs, a penalty of -20 is applied at the terminal state of the trajectory. This large negative reward emphasizes the critical nature of mortality when evaluating treatment effectiveness.
- **Clinical Improvement Rewards.** Additional positive rewards are assigned when key physiological markers improve within prespecified future windows:
  - **Lactate Decrease (+1):** A decrease in lactate levels within the next six hours, indicating improved metabolic status.
  - **Mean Blood Pressure Increase (+1):** MBP rising to ≥ 65 mmHg within the next four hours, for cases where the initial MBP is below this threshold, reflecting improved hemodynamic stability.
  - **SOFA Score Decrease (+3):** A reduction in SOFA score within the next six hours, highlighting improved multi-organ function.
  - **Reduced Norepinephrine Usage (+1):** A decrease in norepinephrine dosage within the next four hours, suggesting better cardiovascular stability.

E. Detailed Experiment Results

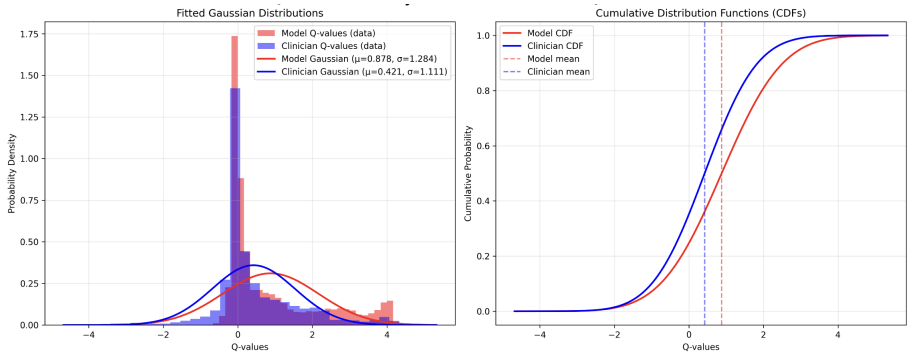
We present detailed comparisons between RL models and baselines.

Figure 7. FQE comparison between the learned patient model and clinician baseline (Dual Stepwise).



Dual Stepwise Model (max_step=0.2): distribution of Q-values from model (red) and clinician (blue).

Figure 8. FQE comparison between the learned patient model and clinician baseline (LSTM BD).



LSTM BD (10 bins): distribution of Q-values from model (red) and clinician (blue).

F. Model Implementation and Training

F.1. Double Q-learning

We use double-Q learning for all RL models in this paper. For our simplified version of the double-Q learning for stability, we adopt two Q-networks, and when Q-value is needed, the minimum scalar output across the two network is taken. Similarly, we adopt two target networks and take the minimum across them when computing the temporal difference (TD) error.

F.2. Binary vp1 Model

The Binary vp_1 model represents the simplest action space, treating vasopressin administration as a binary decision (on/off). The vp_2 levels are used as part of the state. Training employs standard Q-learning TD loss with conservative logsumexp regularization over both action choices to prevent overestimation of out-of-distribution actions. The model processes states through fully connected networks with ReLU activations, using double Q-learning with target networks for stability.

F.3. Dual Mixed Model

The Dual Mixed model handles two action dimensions: vp_1 (binary, 0 or 1) and vp_2 (continuous, 0-0.5 mcg/kg/min). This continuous action space allows for fine-grained dosing control. For action selection of vp2, the model samples 50 candidate actions per state during both training and inference, evaluating Q-values for each to select optimal doses. The conservative Q-learning penalty is computed through importance sampling over randomly sampled actions from the continuous space.

Table 7. Comprehensive comparison of Q-learning models including Binary, Dual Mixed, and Dual Block Discrete (with 3, 5, 10 bins) variants across different conservatism levels (α).

Model	Config	α	vp_1 (%)	Q/step	**ΔQ/step**	vp_1 C. (%)	vp_2 C. (%)
Clinician	–	–	38.8	0.000	0.000	100.0	–
Binary vp_1	–	0.000	77.3	0.792	0.023	51.6	–
Dual Mixed	–	0.000	87.7	1.367	0.177	41.2	–
Dual BD	3 bins	0.000	71.4	1.734	0.123	55.2	57.1
Dual BD	5 bins	0.000	96.3	2.383	0.265	39.7	26.4
Dual BD	10 bins	0.000	92.3	4.673	**0.309**	42.4	13.3
Binary vp_1	–	0.001	79.8	0.869	0.028	44.5	–
Dual Mixed	–	0.001	69.4	1.302	0.184	56.6	–
Dual BD	3 bins	0.001	82.2	1.614	0.110	46.9	55.5
Dual BD	5 bins	0.001	93.6	2.281	0.217	41.3	27.2
Dual BD	10 bins	0.001	96.6	4.445	0.292	39.1	14.1
Binary vp_1	–	0.010	66.2	0.774	0.024	58.7	–
Dual Mixed	–	0.010	77.8	1.345	0.185	52.5	–
Dual BD	3 bins	0.010	76.8	1.576	0.127	53.1	56.3
Dual BD	5 bins	0.010	84.9	2.091	0.119	43.9	40.3
Dual BD	10 bins	0.010	92.8	4.072	0.245	41.4	16.6

F.4. Dual Block Discrete Model

The Dual Block Discrete model discretizes the continuous vp_2 action space into configurable bins while maintaining vp_1 as binary, creating a factored discrete action space with $2 \times N$ (combined with the binary vp_1 action) total actions where N is the number of vp_2 bins. The discretization uses uniform binning from 0 to 0.5 (mcg/kg/min) with bin centers used for continuous dose reconstruction. Q-values are computed for all discrete action combinations using one-hot encoding, enabling efficient batch evaluation. The conservative Q-learning penalty applies logsumexp over all valid discrete actions to maintain conservative value estimates.

F.5. Dual Stepwise Model

The Dual Stepwise model implements an incremental dosing approach where vp_2 changes are limited to fixed steps (e.g., single step ± 0.05, max step± 0.1 mcg/kg/min) from the current dose. This creates a context-dependent action space that respects clinical constraints on dose adjustments. The model augments state representations with one-hot encoded current vp_2 dose bins and maintains valid action masks to prevent out-of-bounds doses. The action space consists of vp_1 (binary) $\times$ vp_2 changes, with conservative Q-learning penalties computed only over valid actions given the current dose context.

F.6. LSTM Block Discrete Model

The LSTM Block Discrete model extends the block discrete approach with sequential modeling capabilities. It processes patient trajectories through LSTM layers to capture temporal dependencies in treatment responses. The model uses sequence lengths of 5 timesteps with 2-step burn-in periods for hidden state initialization. Training employs a specialized replay buffer that maintains overlapping sequences with mortality-weighted sampling priorities. The architecture combines LSTM encoders (2 layers, 32 hidden units) with Q-networks that process concatenated LSTM outputs and state features. Actions are discretized into 10 norepinephrine dosing levels, with conservative Q-learning penalties computed over the discrete action space.

G. Model Architectures

The model architectures for general Q-learning networks is shown in Table 9. The model architectures for LSTM Block Discrete networks is shown in Table 10.

Table 8. Comparison of Binary vp_1 and Dual Stepwise models with vasopressor persistence policy at different maximum step sizes.

Model	α	vp_1 (%)	Q/step	ΔQ/step	vp_1 C. (%)	vp_2 C. (%)
Clinician	–	38.8	0.000	0.000	100.0	–
Binary vp_1	0.001	79.8	0.869	0.028	44.5	–
Dual Mixed	0.010	77.8	1.345	0.185	52.5	–
max_step = 0.1 (mcg/kg/min)						
Dual Stepwise	0.000000	34.6	0.046	0.103	61.1	17.9
Dual Stepwise	0.000100	12.4	0.094	0.136	65.1	17.7
Dual Stepwise	0.001000	82.9	0.252	0.128	47.4	24.5
Dual Stepwise	0.010000	45.9	-0.337	0.114	60.7	22.3
max_step = 0.2 (mcg/kg/min)						
Dual Stepwise	0.000000	97.3	4.021	**0.511**	38.1	11.7
Dual Stepwise	0.000100	31.3	0.134	0.143	70.8	16.4
Dual Stepwise	0.001000	62.0	0.177	0.130	61.2	24.1
Dual Stepwise	0.010000	38.7	-0.184	0.254	64.6	13.5

Table 9. Q Network Architecture (Binary/Mixed/Block/Stepwise)

Component	**Architecture Details**
Input Layer	State (`state_dim`) + Action (`action_dim`)
Hidden Layers	[State, Action] $\xrightarrow{\text{FC (ReLU)}}$ 128 $\xrightarrow{\text{FC (ReLU)}}$ 128 $\xrightarrow{\text{FC (ReLU)}}$ 64
Output Layer	64 $\xrightarrow{\text{FC}}$ 1 (Q-value)
Initialization	Xavier Uniform
Function	$Q(s, a) \to \mathbb{R}$
Networks	Dual Q-networks (Q1, Q2) with dual target networks

H. Hyper-parameter Configurations

The shared hyper-parameters are shown in Table 11. The hyper-parameters specific to different action space models are shown in Table 12.

Table 10. LSTM Q Network Architecture (LSTM Block Discrete CQL)

Component	**Architecture Details**
Input	State (`state_dim`)
Feature Extractor	[State] $\xrightarrow{\text{FC (ReLU)}}$ 32 $\xrightarrow{\text{Dropout(0.1)}}$ $\xrightarrow{\text{FC (ReLU)}}$ 32 $\xrightarrow{\text{Dropout(0.1)}}$
LSTM Layers	32 $\xrightarrow{\text{LSTM}}$ `2 layers` $\times$ 32 units (with dropout=0.1)
Output Layer (Q-head)	32 $\xrightarrow{\text{FC}}$ `num_actions` (Q-values)
Initialization	Xavier Uniform
Function	$Q(s, a) \rightarrow \mathbb{R}^{\text{num\_actions}}$
Networks	Dual Q-networks with LSTM, dual target networks

Table 11. Common Hyperparameters Across All Models

Hyperparameter	**Value**	**Description**
Learning Rate (lr)	10^{-3}	Adam optimizer learning rate
Batch Size	128	Training batch size
Gamma (γ)	0.95	Discount factor for future rewards
Tau (τ)	0.8	Soft target network update rate
Gradient Clipping	1.0	Maximum gradient norm
Epochs	100	Number of training epochs
Validation Batches	10	Batches used for validation
Random Seed	42	Ensures reproducibility

Table 12. Model Configurations: Action Spaces and Hyperparameters

Action Space		**Key Parameters**	
Model	**Configuration**	**Model**	**Hyperparameters**
Binary	1D binary (0/1)	Binary	—
Dual Mixed	2D vp_1 binary, vp_2 continuous vp_1: {0,1}, vp_2: [0,0.5]	Dual Mixed	`num_samples`: 50 (continuous vp_2 selection)
Block Discrete	2×N discrete vp_1: binary, vp_2: N bins	Block Discrete	—
Stepwise	2×M discrete vp_1: binary, vp_2: M steps	Stepwise	`max_step`: {0.1, 0.2} `step_size`: 0.05
LSTM Block	N discrete (vp_2 only)	LSTM Block	`seq_len`: 5, `burn_in`: 2