Guardian-regularized Safe Offline Reinforcement Learning for Smart Weaning of Mechanical Circulatory Devices

Aysin Tumay* ATMAY@UCSD.EDU
University of California, San Diego

Sophia Sun* SHS066@UCSD.EDU
University of California, San Diego

Sonia Fereidooni SFEREIDOONI@UCSD.EDU
University of California, San Diego

Aaron Dumas ADUMAS@CALTECH.EDU
California Institute of Technology

Elise Jortberg JORTBERG.E@GMAIL.COM
Abiomed

Rose Yu ROSEYU@UCSD.EDU
University of California, San Diego

Abstract

We study the sequential decision-making problem for automated weaning of mechanical circulatory support (MCS) devices in cardiogenic shock patients. MCS devices are percutaneous micro-axial flow pumps that provide left ventricular unloading and forward blood flow, but current weaning strategies vary significantly across care teams and lack data-driven approaches. Offline reinforcement learning (RL) has proven to be successful in sequential decision-making tasks, but our setting presents challenges for training and evaluating traditional offline RL methods: prohibition of online patient interaction, highly uncertain circulatory dynamics due to concurrent treatments, and limited data availability. We developed an end-to-end machine learning framework with two key contributions (1) **C**linically-aware **O**OD-**r**egularized **M**odel-based **P**olicy **O**ptimization (CORMPO) a density-regularized offline RL algorithm for out-of-distribution suppression that also incorporates clinically-informed reward shaping and (2) a Transformer-based probabilistic digital twin that models MCS circulatory dynamics for policy evaluation with rich physiological and clinical metrics. We prove that CORMPO achieves theoretical performance guarantees under mild assumptions. CORMPO attains a higher reward than the offline RL baselines by 28% and higher scores in clinical metrics by 82.6% on real and synthetic datasets. Our approach offers a principled framework for safe offline policy learning in high-stakes medical applications where domain expertise and safety constraints are essential.

Data and Code Availability. The code repository[1] includes the full implementation of our method and the synthetic dataset. The real-world dataset contains human-subject information and cannot be publicly released.

Institutional Review Board (IRB). This study used real patient data under IRB approval.

1. Introduction

MCS devices assist the heart by pumping oxygen-rich blood from the left ventricle into the ascending aorta, supporting patients with compromised cardiac function. Weaning from MCS is a series of flow controls over a period of time in which the clinician aims to reduce flow support while maintaining stable hemodynamics, prior to explanting the MCS device (Atti et al., 2022). Reducing the pump flow level (P-Level) is entirely at the discretion of the clinician: the manufacturer's instructions for use suggest reducing by levels of 2, and evaluating at each reduction for evidence of deterioration. However, constant monitoring

* These authors contributed equally

1. available at https://github.com/Rose-STL-Lab/CORMPO

of the patient state is heuristics-based without any empirical or theoretical basis.

Deep reinforcement learning (RL) has shown great promise in automating sequential decision making in medical treatments, with works exploring clinical conditions such as sepsis (Raghu et al., 2017; Komorowski et al., 2018) and cancer (Tseng et al., 2017; Eckardt et al., 2021). With RL's ability to learn sequential decisions from real-world datasets, a data-driven policy can reduce clinician decision fatigue and offer richer guidance compared to rule-based guidelines.

Our application acknowledges three challenges. First, medical treatments cannot be learned via online interaction or exploration on patients, nor can the learned policy be directly evaluated in the real world. Second, the dynamics of weaning MCS devices are highly uncertain and require clinical discretion: patients on MCS are likely also receiving other sources of treatment such as surgery and medications, which the devices's learning algorithm does not have access to. Due to these challenges, state-of-the-art safe offline reinforcement learning algorithms such as uncertainty penalization (Yu et al., 2020) or value regularization (Ma and Kallus, 2022) becomes over-conservative and unstable in our setting.

We propose an end-to-end pipeline for learning clinically informed MCS weaning strategies. Our method, **C**linically-aware **O**OD-**r**egularized **M**odel-based **P**olicy **O**ptimization (CORMPO) tackles the challenges by leveraging a probabilistic digital twin for evaluation, incorporating domain-specific metrics to encourage medically salient behaviors, and developing a density-regularized offline RL algorithm to ensure policy safety while optimizing performance. We show both theoretical and empirical support for our method's strong performance. In summary, our contributions are:

1. We present a Markov Decision Process (MDP) formulation for learning MCS weaning with offline RL models. To evaluate the models, we develop domain-specific clinically-aware metrics and a transformer-based probabilistic digital twin that models MCS circulatory dynamics.

2. Our offline RL algorithm CORMPO utilizes reward shaping and a novel density-based OOD (out-of-distribution) penalization method. The algorithm achieves performance guarantees under mild assumptions, and outperforms offline RL baselines by 28% in physiological reward and by 82.6% in clinical metrics on real and synthetic datasets.

While CORMPO is developed for the specific application of MCS weaning, our methodological contributions, including the density-based OOD safeguarding and clinically-informed reward design, are broadly applicable to many medical decision-making applications that face similar challenges as ours.

2. Related Work

Safety-aware Offline RL. Early safety-aware offline RL methods (Liu et al., 2021) use a dual-policy framework to transform potentially unsafe actions, while Yang et al. (2022) addressed distribution shifts by regularizing the stationary state–action distribution of the current policy to match that of the offline dataset. Ran et al. (2023) proposed Policy Regularization with Dataset Constraint which explores near-dataset state–action pairs using a nearest-neighbor search. The single-neighbor selection may induce bias, a limitation we mitigate by predicting continuous probabilities for each state–action pair. Ma and Kallus (2022) offers a simple and effective means of directly regularizing Q-value estimates by penalizing values outside the support of the behavior policy, and Wu et al. (2022) pursues the same goal using density estimates of state–action pairs. Wang et al. (2023) further advances this idea by employing a FlowGAN-based model for density estimation. However, these safety-focused methods struggle to generalize to real-world datasets, where small data, high dimensionality, and environment stochasticity undermines their effectiveness and stability.

Learning for safe medical decision making. Prasad et al. (2018) pioneered the application of RL for weaning mechanical ventilation, yet their fitted Q-iteration approach struggled with suboptimal clinical data. Tang et al. (2018) leveraged recurrent neural networks to capture temporal dependencies for dynamic treatment recommendations, although imitation learning limits performance to clinician-level decisions. Kuang et al. (2024) built patient-specific cardiac hemodynamic digital twins via physics-informed self supervised learning. Lingsch et al. (2024) proposed neural surrogates for PDE forward simulation and inverse parameter estimation on simulated data. Most similar to our work, Yan et al. (2025) utilizes a classifier to construct safety constraints by a OOD data classifier for offline RL for Sepsis treatment. In contrast

to these methods, our work proposes a novel density-based clinically-aware offline learning algorithm, and presents an end-to-end machine learning framework - including a probabilistic digital twin for evaluation, and designing domain-specific medical metrics.

3. Background and Formulation

Offline Reinforcement Learning. In this work, we formulate our setting as a Markov decision process (MDP), defined by the tuple $M = (\mathcal{S}, \mathcal{A}, T, r, \mu_0, \gamma)$, with state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, reward function $r(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, initial state distribution μ_0, and discount factor, $\gamma \in (0, 1)$. Reinforcement Learning algorithms aim to find a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected cumulative reward $\mathbb{E}_{\pi, s_0 \sim \mu_0} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$. The optimal policy is defined as,

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi, s_0 \sim \mu_0} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]. \quad (1)$$

The *Offline RL setting* is when the algorithm only has access to a dataset sampled from the environment $\mathcal{D}_{\text{env}} = \{(s_i, a_i, r_i, s'_i)\}_i$ under a behavior policy π^{B} but cannot interact with the environment.

Mechanical Circulatory Support (MCS). Left sided forward flow MCS devices are medical devices designed to assist the heart in pumping blood from the left ventricle into the ascending aorta to deliver oxygenated blood to the body. Cardiogenic Shock (CGS) is a syndrome characterized by cardiac output insufficient for end organ perfusion. Hemodynamically, patients in CGS exhibit low systolic blood pressures, low mean aortic blood pressures, and high heart rates. CGS's mortality rate is historically 50-80% (Vahdatpour et al., 2019; Sieweke et al., 2020). For patients in severe CGS, MCS plays an integral role in improving blood pressure, maintaining organ perfusion, and aiding heart muscle recovery. As the patient shows signs of improvement, the care team begins to wean the patient from MCS support. The weaning process includes step-wise reduction in MCS performance P-Level with regular assessment of patient response, see Figure 7 for examples.

In order to learn and evaluate weaning strategies, we need an environment that predicts the patient response to the proposed change in P-level and evaluates the quality of that P-level choice. Although there exist some physics-informed models and numerical simulators (Kuang et al., 2024; Lingsch et al., 2024; Burkhoff et al.) of patient hemodynamics, they are often deterministic and not suitable for long time-horizon simulation. Existing solutions fail to account for noise in the real-life patient data and partial observation, due to unobserved treatments (e.g., surgery, medications) and per-patient variability. Figure 1 showcases one of the challenges: data is sparse for "bad" cases, where the pump level is low and the patient is unstable (red shaded area). This sparsity results in high epistemic uncertainty in this state-action region. To realistically evaluate a weaning strategy, a digital twin model that can probabilistically quantify uncertainty over a significant time-horizon is integral.

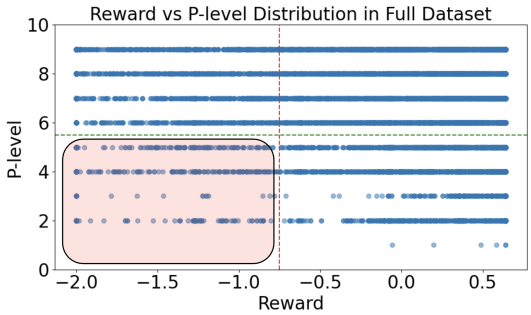


Figure 1: Illustration of data sparsity in low reward, low P-Level region (shaded in red).

4. Methodology

In this section, we detail all components of our algorithm (see Appendix C for pseudocode).

MDP Design for MCS. We first formulate the MCS weaning problem as an MDP. The challenge in formulating the environment is balancing the rich information of medical time series with the learning challenges of a high-dimensional Markov Decision Process (MDP). We define each *state* in the MDP to consist of t time-steps of k different physiological features, i.e. $\mathcal{S} \subseteq \mathbb{R}^{txk}$. The *action* space is $\mathcal{A} = \{2, 3, \cdots, 9\}$, corresponding to pump level P2 to P9 on the MCS device. The objective is to optimize patient outcome with a clinically appropriate weaning strategy. For the offline RL problem, we organize the patient data into a replay buffer dataset of $\mathcal{D} = \{(s_i, a_i, s'_i, r_i)\}_i$ according to the formulation. The state space, action space, reward, and MDP design is informed by expert recommendation and empirical results as presented in Appendix B. Under the MDP formulation, we will then describe our digital twin for evaluation, followed by the setup for offline RL training.

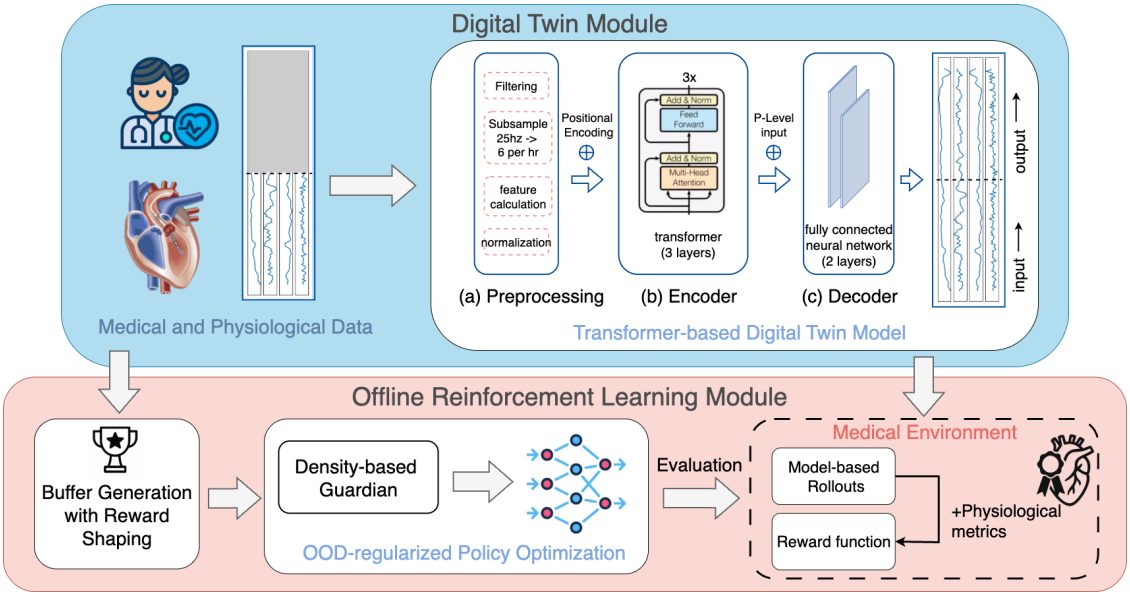


Figure 2: System diagram of the proposed framework. **Digital twin module:** We use a transformer encoder to learn a latent representation of the patient's history, concatenate the representation with the P-Level input, and then decode the output using a fully connected neural network. **Offline RL module with CORMPO:** The replay buffer is created from data with clinical guided reward shaping. We learn a density-based guardian model on the data, whose OOD penalty terms are incorporated during policy training. The learned policies are evaluated in the digital twin-supported medical environment with rich medical metrics.

4.1. Transformer-based Digital Twin.

To simulate patient trajectories during weaning, we develop a Transformer-based digital twin (TDT) that models patient hemodynamic signals under MCS. The digital twin is denoted as $\hat{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, which serves as a proxy of the stochastic transition function for the RL task, i.e. $\hat{T}(s, a) = p(s'|s, a)$. At each timestep, the digital twin receives the current patient state, along with the control action (for the following time step). The digital twin forecasts the next physiological state, enabling safe synthetic "what-if" scenarios by simulating patient responses to candidate weaning actions.

The digital twin's model architecture is shown in Figure 2. The encoder with three multi-head self-attention layers captures temporal dependencies in the multivariate physiological time series. The action is then concatenated to the latent representation and passed to the decoder. The decoder (2 fully connected perceptron layers) predicts the next state. Probabilistic prediction is achieved by retaining dropout ($p = 0.1$) in the decoder layers. We train the model to minimize the MSE between the predicted and observed future states, using historical data.

4.2. Offline Reinforcement Learning.

We propose an algorithm named **C**linically-aware **O**OD-regularized **M**odel-based **P**olicy **O**ptimization (CORMPO) for learning clinically informed MCS weaning strategies. Our algorithmic contribution are in two-fold: (1) we designed task-specific reward shaping to incorporate clinical guidelines into the optimization process, and (2) we introduce a density-based OOD-data suppression algorithm to tackle the safety problem in offline reinforcement learning.

4.2.1. Clinical metrics and reward shaping

Physiological reward reflects well-being from mean arterial pressure (MAP), heart rate, and pulsatility over the past hour. As we not only value high physiological reward but also gradual P-level changes that lead to stable weaning, we shape the physiological reward with Action Change Penalty (ACP) (similar to Yan et al. (2025)), and Weaning Score (WS). ACP accumulates the magnitude of P-level changes over an episode of length T as

$$\text{ACP} = \sum_{i=1}^{T} ||a_{i-1} - a_i||_2, \text{ if } ||a_{i-1} - a_i||_2 > 2.$$

WS rewards the decrease and penalizes the increase in P-level conditioned on the hemodynamic stability of the patient state as follows:

$$\mathrm{WS} = \frac{\sum_{i=1}^{T} \mathbb{I}(\texttt{Is\_Stable}(i), 1) \cdot \texttt{Weaned}(i)}{\sum_{i=1}^{T} \mathbb{I}(\texttt{Is\_Stable}(i), 1)},$$

The definitions for `Is_Stable` and `Weaned` are in Appendix A. Finally, we formulate our reward as

$$r(\cdot) = r_{\text{phys}}(\cdot) - \lambda_1 ACP(\cdot) + \lambda_2 WS(\cdot) \tag{2}$$

where λ_1, and λ_2 are hyperparameters determining the magnitude of shaping of each medical metric.

4.2.2. Density-based guardian for ORL

Our goal is to learn a policy π that maximizes the expected return in the true MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, T, r, \mu_0, \gamma)$, where T denotes the true dynamics, r the reward function, μ_0 the initial state distribution, and $\gamma \in (0, 1)$ the discount factor.

Following the model-based approach, we learn a digital twin $\hat{T}$ from $\mathcal{D}_{\text{env}}$, defining the model MDP as $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{T}, r, \mu_0, \gamma)$. Let

$$\rho^{\pi}_{\hat{T}}(s, a) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t P^{\pi}_{\hat{T},t}(s) \tag{3}$$

denote the discounted occupancy measure under policy π and dynamics $\hat{T}$, where $P^{\pi}_{\hat{T},t}(s)$ is the probability of visiting state s at time t.

We denote the value function under the true dynamics as $V^{\pi}_{\mathcal{M}}(s)$ and the expected return as

$$\eta_{\mathcal{M}}(\pi) = \mathbb{E}_{s_0 \sim \mu_0}[V^{\pi}_{\mathcal{M}}(s_0)].$$

The challenge with model-based offline RL is that learned dynamics models inevitably exhibit varying degrees of accuracy across the state-action space, with errors compounding over multi-step rollouts. While the optimal policy under perfect dynamics may venture beyond the behavioral distribution to achieve higher returns, model inaccuracies in these out-of-distribution (OOD) regions can lead to catastrophic failures, a critical concern in high-stake medical applications. Existing uncertainty-based methods (Yu et al., 2020; Kidambi et al., 2020) conflates two distinct sources of uncertainty: *aleatoric* uncertainty arising from inherent environment stochasticity, and *epistemic* uncertainty stemming from limited data coverage. This conflation leads to over-penalization of in-distribution (ID) states in noisy environments, unnecessarily constraining the policy in well-understood regions of the state space. Such conservative behavior may prevent the policy from executing known-safe actions that are crucial for task completion.

As the density of training data determines model reliability, we propose distinguishing ID and OOD states by directly measuring data support through density estimation, which quantifies epistemic uncertainty independent of environment noise. The density-based safeguard enables CORMPO to selectively penalize OOD states and actions while preserving optimal behavior within the data support.

Density Estimation and Safeguard. We use kernel density estimation (KDE) in implementation. Given the dataset $\mathcal{D}_{\text{env}}$, the KDE estimator is:

$$p_{\text{KDE}}(s, a) = \frac{1}{N} \sum_{i=1}^{N} K_h\left((s, a) - (s_i, a_i)\right) \tag{4}$$

where K_h is a kernel function with bandwidth h, and N is the number of neighbors closest to each (s_i, a_i).

We define the density regularizer as:

$$u(s, a) = \tau - \log(p_{\text{KDE}}(s, a)) \tag{5}$$

where τ is the density threshold of in-distribution data. We detail our process of choosing τ as a percentile in Appendix F.1. The density regularizer is then decomposed into $u_+(s, a) = \max\{u(s, a), 0\}$ and $u_-(s, a) = \min\{u(s, a), 0\}$. Now define the regularized MDP as $\tilde{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{T}, \tilde{r}, \mu_0, \gamma)$ with:

$$\tilde{r}(s, a) = r(s, a) - \lambda u(s, a) \tag{6}$$

where $\lambda = \gamma c \cdot C_{\hat{T}}$, the symbols will be introduced in the next section.

The optimal policy for the density-penalized MDP is obtained by solving:

$$\hat{\pi} = \arg\max_{\pi} \eta_{\tilde{\mathcal{M}}}(\pi). \tag{7}$$

4.2.3. Theoretical Results

We show theoretical support that $\hat{\pi}$ as learned by CORMPO has guaranteed performance in the real MDP $\mathcal{M}$, and achieves near-optimal performance among policies that maintain low penalty under the learned dynamics.

Our guarantee relies on two assumptions. Assumption 1 on bounded rewards is standard in RL theory and holds for most practical applications; assumption 2 captures epistemic uncertainty in supervised learning where model accuracy degrades as we move away from the training distribution.

Assumption 1 (Bounded Rewards, Density Regularizer, and Value Functions)**.** *The reward function is bounded:* $|r(s,a)| \leq r_{\max}$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$*. Consequently,* $V^{\pi}_{\mathcal{M}} \in c\mathcal{F}$ *where* $\mathcal{F} = \{f : \|f\|_{\infty} \leq 1\}$*,* $c = r_{\max}/(1-\gamma)$*. Let* $u : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *be bounded and write its negative part as* $u_{-}(s,a) = \min\{u(s,a), 0\} \leq 0$ *with* $\|u_{-}\|_{\infty} := \sup_{(s,a)} |u_{-}(s,a)| < \infty$*.*

Assumption 2 (Density-Dependent Model Error)**.** *There exists a constant* $C_{\hat{T}} > 0$ *such that for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$*:*

$$d_{\mathcal{F}}(\hat{T}(s,a), T(s,a)) \leq C_{\hat{T}} \cdot u_{+}(s,a) + \epsilon_{approx} \quad (8)$$

where $d_{\mathcal{F}}$ *is the integral probability metric w.r.t.* $\mathcal{F}$*, and* $\epsilon_{approx} > 0$ *represents an irreducible approximation error.*

The constant $C_{\hat{T}}$ depends on the model class capacity and the smoothness of the true dynamics, which we further discuss in Appendix D.

Next, define $|G^{\pi}_{\hat{\mathcal{M}}}(s,a)|$ as the difference between the expected value of transition functions T and $\hat{T}$. By the Telescoping Lemma in Appendix D, we have

$$\begin{aligned} |G^{\pi}_{\hat{\mathcal{M}}}(s,a)| &= \left| \mathbb{E}_{s' \sim \hat{T}(s,a)}[V^{\pi}_{\mathcal{M}}(s')] - \mathbb{E}_{s' \sim T(s,a)}[V^{\pi}_{\mathcal{M}}(s')] \right| \\ &\leq c \cdot d_{\mathcal{F}}(\hat{T}(s,a), T(s,a)) \qquad \text{(Lemma 1)} \\ &\leq c \cdot C_{\hat{T}} \cdot (u_{+}(s,a) + \epsilon_{approx}) \quad \text{(Asm. 2)} \end{aligned}$$

where c is specified as in assumption 1. Bounding value differences allows us to achieve the value bound and optimality results as in uncertainty penalization-based methods (Yu et al., 2020):

Theorem 1 (Conservative Value Bound)**.** *Under Assumptions 1-2, for any policy* π*:*

$$\eta_{\mathcal{M}}(\pi) \geq \eta_{\bar{\mathcal{M}}}(\pi) - \frac{\gamma c \epsilon_{approx} + \beta}{1-\gamma} \quad (9)$$

where $\beta = \lambda \mathbb{E}_{(s,a) \sim \rho^{\pi}_{\hat{T}}}[|u_{-}(s,a)|]$*.*

Theorem 2 (Optimality Gap)**.** *Let* π^{*} *be the optimal policy for the true MDP* $\mathcal{M}$ *and* $\hat{\pi}$ *be the solution to Equation (7). Define* $\delta_{\min} = \min_{\pi} \mathbb{E}_{(s,a) \sim \rho^{\pi}_{\hat{T}}}[u_{+}(s,a)]$ *and* $||u_{-}||_{\infty} = \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |u_{-}(s,a)|$*. Then:*

$$\begin{aligned} \eta_{\mathcal{M}}(\hat{\pi}) \geq \max_{\pi : \mathbb{E}_{\rho^{\pi}_{\hat{T}}}[u_{+}] \leq \delta} \eta_{\mathcal{M}}(\pi) - 2\lambda\delta - \frac{\gamma c\, \epsilon_{approx}}{1-\gamma} \\ - \frac{\lambda}{1-\gamma} \|u_{-}\|_{\infty} \end{aligned} \quad (10)$$

for any $\delta \geq \delta_{\min}$*.*

This result shows that $\hat{\pi}$ achieves near-optimal performance among policies that maintain high density under the learned dynamics. The term $2\lambda\delta$ is the price of conservativeness: the smaller the density budget δ, the tighter the guarantee but the less exploration. Third term on the RHS reflects irreducible model error in the learned dynamics. $||u_{-}||_{\infty}$ in the last term indicates the maximum value of high-density bonus. So, the last term upper-bounds any optimism induced by rewarding high-density areas. Intuitively, our density-based penalty implements an adaptive exploration-exploitation trade-off: In high-density regions where $\log(p_{\text{KDE}}(s,a)) \geq \tau$, the term boosts rewards with a positive bonus, allowing the policy to fully exploit the learned model, and in low-density regions where $\log(p_{\text{KDE}}(s,a))$ is close to zero, the penalty discourages exploration into unsafe areas.

5. Experiments

5.1. Experiment Setup

In this section, we start by introducing the details of our real-life and synthetic datasets. Then, we present results of digital twin learning and the policy evaluation process. Lastly, we compare our proposed model against state-of-the-art offline RL algorithms with qualitative and quantitative results.

Dataset. Our real-life MCS dataset includes 379 patients, with an average length of record of 65.5 hours. We split the patients by ratio 65-15-20 into training, validation, and testing sets. Our clinical data includes 12 features recorded directly or derived from from signals of the MCS device, namely: Mean aortic pressure (MAP), mean pump speed, mean motor current, mean pump flow, left Ventricular Pressure (LVP), left ventricular end diastolic pressure (LVEDP), heart rate (HR), Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Pulsatility, Relaxation Constant (Tau_LV), and elastance estimation (ESE_LV). We downsample the original signal of 25 Hz into 0.00167 Hz (1 sample per 10 minutes) and extract samples with a sliding window of 1 hour.

Digital Twin. Adopting model-based offline policy evaluation (OPE), we utilize our TDT as a surrogate of the real-life MCS environment. We demonstrate that our TDT successfully captures the underlying data dynamics and associated uncertainty in Appendix E, outperforming baselines across accuracy and uncertainty calibration metrics by more than 35%. Our

Metric	Expert	BC	MBPO	MOPO	SVR	CORMPO
Phys. Reward (↑)	1.167	0.101 ± 0.154	1.108 ± 0.028	1.059 ± 0.113	0.278 ± 0.158	$\mathbf{1.224 \pm 0.105}$
ACP (↓)	9.26	3.703 ± 0.141	1.963 ± 0.078	1.907 ± 0.069	2.945 ± 0.107	$\mathbf{0.285 \pm 0.035}$
WS (↑)	-0.091	-0.023 ± 0.008	-0.076 ± 0.004	-0.021 ± 0.010	-0.043 ± 0.007	$\mathbf{0.040 \pm 0.010}$

Table 1: Trained on the noiseless synthetic dataset, we compare CORMPO against different offline RL models in terms of physiological reward, ACP, and WS (↑: higher is better; ↓: lower is better). Evaluation is done on 1000 episodes and averaged over 5 seeds. CORMPO significantly outperforms baselines on physiological reward (14.9%), ACP (82.6%), and WS (2.9 times larger than MOPO).

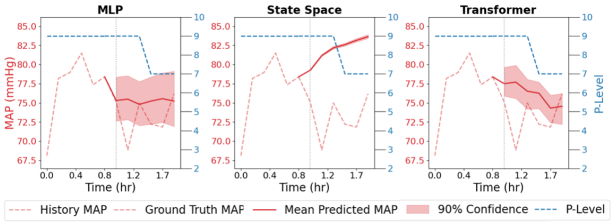


Figure 3: An example of TDT prediction vs. baselines. The Transformer model is more accurate in reflecting response to P-level change and more expressive when capturing large changes in patient state, resulting in its higher accuracy as shown in Table 7.

Transformer model is more accurate in reflecting response to P-level change and more expressive when capturing large changes in patient states, resulting in its higher accuracy.

Synthetic Data Generation. We start by constructing a synthetic dataset for a more controlled study of the robustness and efficacy of our algorithm. The synthetic dataset consists of 5000 trajectory roll-outs from the TDT, starting from a random state in the real dataset and ending at $T = 6$, with an expert policy trained with online RL algorithm SAC (Haarnoja et al., 2018). To mimic the noise setting of our real-life MCS data, we also created a noisy version of the synthetic dataset, by adding a Gaussian noise of $\mathcal{N}(0, 0.2)$ to 80 % of state transitions.

Baselines and metrics. We compare our proposed method against 4 state-of-the-art Offline RL baselines and the expert policy. For the synthetic experiment, expert is a policy learned in the TDT environment; for the real data experiment, expert is doctor's ground truth actions evaluated with TDT roll-outs. `Behavioral Cloning (BC)` is a supervised learning baseline trained without exploration. `Model-Based Policy Optimization (MBPO)` (Janner et al., 2019) fits a dynamics model and then optimizes the policy on real data augmented with short-horizon rollouts from the dynamics model. `Model-based Offline Policy Optimization (MOPO)` (Yu et al., 2020) penalizes the reward using transition uncertainty to encourage conservative policy learning. `Support Value Regularization (SVR)` (Ma and Kallus, 2022) incorporates out-of-distribution (OOD) value regularization to guide policy learning in safe regions. We evaluate each policy for 1000 episodes of $T = 6$, i.e. a horizon of 6-hour and compare performances with respect to physiological reward, Action Change Penalty (ACP), and Weaning Score (WS).

5.2. Synthetic Data Results

We evaluate the performance and noise robustness of CORMPO in the synthetic setting. We present results of learning on the noisy synthetic dataset, deferring implementation details and hyperparameters to Appendix F.1.

In Table 1, we highlight that our method outperforms baselines in reward and ACP, showcasing the functionality of our reward shaping and MDP design. We demonstrate moderate performance in WS by outperforming all baselines except for SVR.

Metric	Expert	MBPO	MOPO	CORMPO
Reward Noiseless	1.16	1.108 ± 0.028	1.059 ± 0.113	$\mathbf{1.224 \pm 0.105}$
Reward Noisy	1.110	1.064 ± 0.031	0.947 ± 0.094	$\mathbf{1.223 \pm 0.106}$
% of Drop	5.74	3.97	10.6	**0.082**

Table 2: Comparison of physiological reward under noiseless and noisy settings, and corresponding percentage drop. Evaluation is done on 1000 episodes and averaged over 5 seeds in the noiseless environment setting. CORMPO shows the lowest degradation under noise, indicating robustness.

In Table 2, we emphasize the robustness of CORMPO in reward. Notably, it shows the smallest reward degradation under noise, outperforming MBPO, and MOPO. Taking the expert policy % of drop in reward as the increase in noise scale, OOD-regularization

Metric	Expert	BC	MBPO	MOPO	SVR	CORMPO
Phys. Reward (↑)	0.557	0.175 ± 0.118	0.420 ± 0.139	0.373 ± 0.129	0.530 ± 0.152	$\mathbf{0.687 \pm 0.106}$
ACP (↓)	1.79	0.068 ± 0.012	0.459 ± 0.032	0.984 ± 0.020	0.599 ± 0.057	$\mathbf{0.018 \pm 0.007}$
WS (↑)	0.053	$\mathbf{0.345 \pm 0.008}$	0.147 ± 0.007	0.042 ± 0.006	0.166 ± 0.003	0.173 ± 0.007

Table 3: Trained on the real-life dataset, we compare CORMPO against different offline RL baselines in terms of physiological reward, ACP, and WS with 1000 episodes averaged over 5 seeds. Evaluation is completed in the noiseless environment setting. ↑: higher is better; ↓: lower is better. CORMPO outperforms the baselines by 28% in physiological reward, and 73% in ACP.

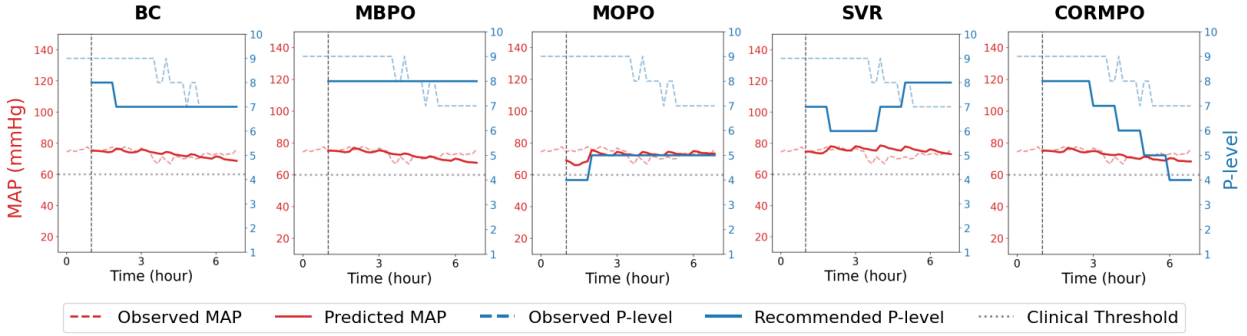


Figure 4: Trained on the real-life dataset, we compare our CORMPO against baselines in 6-hour TDT rollouts. Our TDT predicts stable hemodynamics, away from the clinical threshold, for MAP when guided by CORMPO's optimal policy. CORMPO's WS is 1.0 which is the maximum possible score. While BC policy yields limited weaning, MBPO, MOPO, and SVR acts opposite to the weaning behavior because patient stability suggests gradual decrease in P-level. CORMPO results in the most successful weaning in this sample roll-out.

of CORMPO indicates the functionality of our method. CORMPO also outperforms RL baselines on the noisy synthetic dataset setting.

5.3. Real Dataset Results

We further demonstrate the performance of CORMPO on the real dataset in Figure 4 and Table 3. In Table 3, CORMPO demonstrates superior overall performance, achieving the highest reward while maintaining the lowest ACP and a WS outperforming most baselines, indicating both successful policy optimization and action stability. While BC achieves the highest WS, this is offset by poor reward performance, which can be attributed to the inherent variability in clinician-chosen P-levels and the lack of active exploration leading to over-conservative policy. Specifically, the highest ACP value originating from the clinician P-levels shows the stochasticity of the expert actions, depicted in Figure 5 row 1 column 1 with the expert resulting in the highest ACP. Therefore, BC avoids P-level increases and defaults to conservative behavior, driven by stochasticity in the clinician data. MBPO is the third-best performing model after SVR. The reward of MOPO being worse than MBPO and CORMPO underlines that it over-penalizes the model-based rollouts that diverge into OOD states when the real dataset is already largely stochastic and noisy. SVR also demonstrates

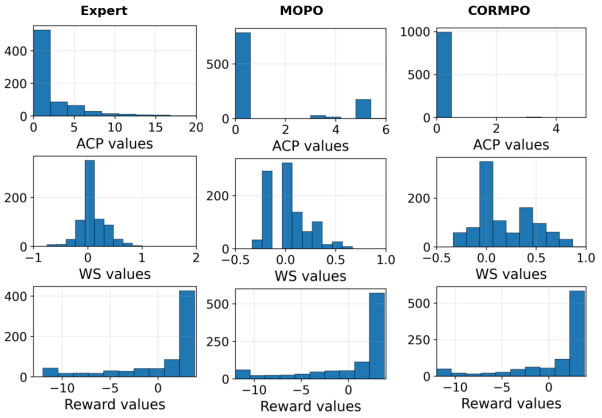


Figure 5: Comparison of physiological reward, ACP, and WS distribution of expert, MOPO, and CORMPO policies. CORMPO suppresses actions with high ACP and results in higher WS compared to baselines, also reducing the high portion of negative rewards in expert policy and achieves higher rewards overall.

	MBPO	MOPO	CORMPO
Elapsed time (s)	1185.670	1165.866	1195.765

Table 4: Elapsed training time of MBPO, MOPO, and CORMPO. Trained for 100 epochs on an NVIDIA A100 80GB GPU with identical hyperparameters; the small difference ($<$20 s) indicates CORMPO adds no significant overhead.

suboptimal policy performance, as reflected by its high ACP. In contrast, CORMPO pushes the policy in the higher reward region without sacrificing WS and ACP as seen in Figure 5. We depict further qualitative and hyperparameter sensitivity analysis of CORMPO in Appendix F.2.

Our KDE implementation is Facebook AI Similarity Search Johnson et al. (2019) which provides a fast approximate nearest-neighbor that has $O(NlogN)$ complexity while traditional KDE's have $O(N^2)$. Therefore, our KDE training takes 0.37 seconds on our train dataset of 13,399 samples on NVIDIA A100 80GB GPU. We save the model checkpoint and use it for inference during policy training. This does not impose a computational overhead in CORMPO as demonstrated in Table 4.

6. Discussion

We presented CORMPO, a comprehensive framework for safe weaning of mechanical circulatory support devices using offline reinforcement learning. Our approach addresses three critical challenges in medical decision-making: the prohibition of online patient interaction, highly uncertain circulatory dynamics, and limited data availability. We propose a safe offline RL framework that incorporates clinical knowledge via reward shaping and uses density-based regularization to avoid OOD regions, evaluated in our Transformer-based digital twin environment. Our theoretical analysis provides performance guarantees under mild assumptions, while experimental validation on synthetic and real patient data demonstrates consistent outperformance of established offline RL baselines across clinically-relevant metrics, including physiological reward, action change penalty, and weaning score. In conclusion, our work presents a complete offline RL methodology for developing a data-driven safe medical decision-making algorithm.

Choice of offline policy evaluation method. In this paper, we chose model-based offline policy evaluation (OPE) (a widely used approach, see Zhang et al. (2021); Voloshin et al. (2021)) over the importance sampling-based approach typically used in medical RL literature (Prasad et al., 2018). The choice is primarily because of two reasons: (1) In the medical setting, it is important to show the "what-if" scenarios resulting from different P-levels to the physicians and decision makers. We developed the digital twin to this end, and use it to evaluate the policies such that it's more interpretable to our medical collaborators. As importance sampling (IS) methods do not provide forecasted trajectories, it makes our system less trustworthy to the physicians. (2) A primary challenge in this problem is sparse data coverage (see Figure 1). This sparsity fundamentally limits the reliability of Importance Sampling (IS) methods for policy evaluation. Since the dataset consists of human expert actions without an explicit behavior policy, any IS-based approach requires first learning an approximation behavior policy from the data. This introduces a compounding of errors: (a) the inherent high variance of IS in low-coverage regions, where importance weights can become arbitrarily large for state-action pairs rarely visited by the expert, and (b) systematic errors from behavior policy mis-estimation, which are most severe precisely in these same low-coverage regions where we have insufficient data to learn $\pi^{behavior}$ accurately. These two reasons make IS-based evaluation unreliable for our setting. Some studies have shown that combining model-based and IS methods result in more robust off-policy evaluation (Voloshin et al., 2021; Thomas and Brunskill, 2016). We defer it to future work to hold a detailed evaluation with the hybrid method.

Limitations and Future Work. As next steps of this work, we will have the clinical metrics used in this work, such as hemodynamic stability, reviewed by intensive care unit doctors for their suitability and actionability. The definitions are author-designed proxies based on existing device guidance. How the reward shaping mechanism affects the information quality of the reward signal should be furthered studied as well. Additionally, the performance sensitivity to hyperparameter selection for reward shaping and density threshold suggests opportunities for improving the robustness of the algorithm. Future work could explore the use of generative density estimators and threshold-free density-based regularization. For evaluation and applicability, we will explore hybrid OPE methods and transferability of learned policies across different patient populations and clinical settings.

Acknowledgment

This work was supported in part by a research grant from Abiomed, Inc., a Johnson & Johnson company, the U.S. Army Research Office under Army-ECASE award W911NF-07-R-0003-03, the U.S. Department Of Energy, Office of Science, IARPA HAYSTAC Program, and NSF Grants #2205093, #2146343, #2134274, CDC-RFA-FT-23-0069, DARPA AIE FoundSci and DARPA YFA.

References

Varunsiri Atti, Mahesh Anantha Narayanan, Brijesh Patel, Sudarshan Balla, Aleem Siddique, Scott Lundgren, and Poonam Velagapudi. A comprehensive review of mechanical circulatory support devices. *Heart International*, 16(1):37–48, March 2022. doi: 10.17925/HI.2022.16.1.37. URL https://doi.org/10.17925/HI.2022.16.1.37.

Edward Buitenwerf, Mats F Boekel, Marieke I van der Velde, Magiel F Voogd, Michiel N Kerstens, Götz JKG Wietasch, and Thomas WL Scheeren. The haemodynamic instability score: Development and internal validation of a new rating method of intra-operative haemodynamic instability. *European Journal of Anaesthesiology| EJA*, 36(4):290–296, 2019.

Daniel Burkhoff, Marc L. Dickstein, and Thomas Schleicher. Harvi-online. https://harvi.online. Accessed November 2025.

Jan-Niklas Eckardt, Karsten Wendt, Martin Bornhaeuser, and Jan Moritz Middeke. Reinforcement learning for precision oncology. *Cancers*, 13(18): 4624, 2021.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102 (477):359–378, 2007.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

James D Hamilton. State-space models. *Handbook of econometrics*, 4:3039–3080, 1994.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. doi: 10.1109/TBDATA.2019.2921572.

R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21810–21823, 2020.

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

Keying Kuang, Frances Dean, Jack B. Jedlicki, David Ouyang, Anthony Philippakis, David Sontag, and Ahmed M. Alaa. Med-real2sim: Non-invasive medical digital twins using physics-informed self-supervised learning. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0b081a44ed0b8c0c4aa6bd886a60bea4-Paper-Conference.pdf.

Alan Li, Zihao Zhou, Elise Jortberg, and Rose Yu. Forecasting aortic pressure cross-cohort with deep sequence models. In *2022 Computing in Cardiology (CinC)*, volume 498, pages 1–4. IEEE, 2022.

Levi E. Lingsch, Dana Grund, Siddhartha Mishra, and Georgios Kissas. Fuse: Fast unified simulation and estimation for PDEs. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/266c0f191b04cbbbe529016d0edc847e-Paper-Conference.pdf.

Y. Liu, X. Luo, P. Zhou, et al. Towards safe reinforcement learning with a safety editor policy. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 17802–17814, 2021.

Xuefeng Ma and Nathan Kallus. Supported value regularization for offline reinforcement learning. In

Advances in Neural Information Processing Systems, volume 35, pages 32878–32890, 2022.

N. Prasad, L. F. Cheng, C. Chivers, M. Draugelis, and B. Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. In *Proceedings of the 3rd Machine Learning for Healthcare Conference (MLHC)*, pages 282–299, 2018.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163. PMLR, 2017.

L. Ran, Y. Zhang, X. Liu, and T. Yuan. Policy regularization with dataset constraint for offline reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Jan-Thorben Sieweke, Dominik Berliner, Jörn Tongers, L Christian Napp, Ulrike Flierl, Florian Zauner, Johann Bauersachs, and Andreas Schäfer. Mortality in patients with cardiogenic shock treated with the impella cp microaxial pump for isolated left ventricular failure. *European Heart Journal: Acute Cardiovascular Care*, 9(2):138–148, 2020.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Sophia Sun, Wenyuan Chen, Zihao Zhou, Sonia Fereidooni, Elise Jortberg, and Rose Yu. Data-driven simulator for mechanical circulatory support with domain adversarial neural process. In *6th Annual Learning for Dynamics & Control Conference*, pages 1513–1525. PMLR, 2024.

X. Tang, Y. Jia, J. Sun, and Y. Fan. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pages 2447–2456, 2018.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International conference on machine learning*, pages 2139–2148. PMLR, 2016.

Huan-Hsin Tseng, Yi Luo, Sunan Cui, Jen-Tzung Chien, Randall K Ten Haken, and Issam El Naqa. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical physics*, 44 (12):6690–6705, 2017.

Cyrus Vahdatpour, David Collins, and Sheldon Goldberg. Cardiogenic shock. *Journal of the American Heart Association*, 8(8):e011991, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Aaron Voelker, Ivana Kajić, and Chris Eliasmith. Legendre memory units: Continuous-time representation in recurrent neural networks. *Advances in neural information processing systems*, 32, 2019.

Cameron Voloshin, Hoang M. Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS Track on Datasets and Benchmarks)*, 2021.

Jianhong Wang, Zhizhou Wang, Jian Zhang, Changjian Xu, Ke Li, and Weinan Zhang. Constrained policy optimization with explicit behavior density for offline reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 36485–36506. PMLR, 2023.

Yifan Wu, Wenxuan Zhou, Chen Bai, Yang Yu, and Hongyuan Zha. Supported policy optimization for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30072–30084, 2022.

Runze Yan, Xun Shen, Akifumi Wachi, Sebastien Gros, Anni Zhao, and Xiao Hu. Offline guarded safe reinforcement learning for medical treatment optimization strategies. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

S. Yang, S. Nair, T. Ma, and C. Finn. Regularizing a model-based policy stationary distribution to stabilize offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 12655–12667, 2022.

T. Yu, S. Kumar, A. Gupta, et al. Mopo: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14129–14142, 2020.

Michael R. Zhang, Tom Le Paine, Ofir Nachum, Cosmin Paduraru, George Tucker, Ziyu Wang, and Mohammad Norouzi. Autoregressive dynamics models for offline policy evaluation and optimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=tK_F6Do0P7z.

Appendix A. Medically-informed Metrics

Action Change Penalty (ACP) Yan et al. (2025): Abrupt and extreme changes in P-level may maximize rewards; however, they can induce physiological instability in a real-world setting. ACP gauges policy volatility and is given by:

$$\text{ACP} = \sum_{i=1}^{T} ||a_{i-1} - a_i||_2,$$

where a_{i-1} is an action at state $i-1$, a_i is a subsequent action, and T is the episode length. Lower ACP values indicate stable physiology and safe weaning, but note that a value of 0 is undesirable as the P-level must be lowered for weaning.

Weaning Score (WS): To capture satisfactory weaning patterns, we support P-level reductions at most every 1 hour when the patient is observed as hemodynamically stable for the past 1 hour as depicted in Eq. 4.2.1. Higher weaning scores signify an appropriate reduction in P-level during relatively stable physiological states. We employ 2 definitions of stability based on (1) clinical safety limits and (2) gradient of the past hemodynamic state. First definition is as follows:

$$\texttt{Is\_Stable}(i) = \text{MAP}(i) > \tau_{\text{MAP}} \wedge \text{HR}(i) > \tau_{\text{HR}} \wedge \text{Pulsat}(i) > \tau_{\text{Pulsat}}$$

where $\tau_{\text{MAP}} = 60, \tau_{\text{HR}} = 50$ and $\tau_{\text{Pulsat}} = 10$, indicating limits of hemodynamic stability (see Table 5 for stability limits). Since our state design represents 1 hour in 10-minute time steps, we calculate the compared MAP value for a state as, $\text{MAP}(i) = \min_{1 \le t \le 6} \text{MAP}(i,t)$, same for HR and pulsatility. T is the episode length and $i = 0$ indicates the initial state. Higher weaning scores denote proper lowering of P-level when at a stable state, and low or negative scores imply that P-level is increased despite having healthy physiological indicators. Second definition follows analytically as,

$$\texttt{Is\_Stable}(i) = \left|\frac{\partial \text{MAP}(i)}{\partial t}\right| < \tau_1 \wedge \left|\frac{\partial \text{HR}(i)}{\partial t}\right| < \tau_2 \wedge \left|\frac{\partial \text{Pulsat}(i)}{\partial t}\right| < \tau_3 \tag{11}$$

where $\tau_{\text{MAP}} = 1.36, \tau_{\text{HR}} = 2.16$ and $\tau_{\text{Pulsat}} = 1.95$, indicating a proxy for stability with a low gradient value of 3 hemodynamic indicators in the past state chosen with statistical significance tests.

Among these two definitions, we utilize the clinical threshold-based stability in the reward shaping while we evaluate the policies with the gradient-based WS definition. Second part of the WS metric is as follows.

$$\texttt{Weaned}(i) = \begin{cases} -1, & \text{if } a_{i+1} - a_i > 0, \\ a_{i+1} - a_i, & \text{if } a_{i+1} - a_i \in \{1, 2\}, \\ 0, & \text{otherwise.} \end{cases}$$

Physiological Reward: The reward generally reflects the well-being of the patient, according to the mean arterial pressure (MAP), heart rate (HR), and pulsatility of the past hour. Our design follows the clinically defined ranges for hemodynamic stability while caring for the smoothness and differentiability of the function.

The reward design in Table 5 is staircase-shaped, which has two drawbacks: non-differentiability and a sparse signal. We reformulate the hemodynamic instability score in the following way.

- **Heart Rate Penalty Function** The heart rate penalty function penalizes deviations from an optimal heart rate of 75 bpm using a quadratic penalty:

$$P_{\text{hr}}(hr) = \text{ReLU}\left(\frac{(hr - 75)^2}{250} - 1\right) \tag{12}$$

  where $\text{ReLU}(x) = \max(0, x)$. This function has zero penalty for heart rates in the range $[50, 100]$ bpm and applies quadratic penalties for heart rates outside this range.

Score Component	Value	Score	
Hemodynamic Variable	MAP	≥ 60	0
	50 to 59	1	
	40 to 49	3	
	< 40	7	
Minimum MAP in window	≥ 60	0	
	50 to 59	1	
	40 to 49	3	
	< 40	7	
Time Spent MAP < 60 mmHg (%)	0	0	
	2	1	
		5	3
	> 5	7	
	Pulsatility	> 20	0
	10-20	5	
	< 10	7	
HR	> 100	3	
	< 50	3	
LVEDP	> 20	7	
	15 to 20	4	
	< 15	3	
CPO	0.6 to 1	1	
	< 0.6	3	
	< 0.5	5	

Table 5: Hemodynamic instability score table from Buitenwerf et al. (2019). We use a modified version of this table as our physiological reward. When used for evaluating the learned policy as a reward function, we multiply the risk score by -1.

- **Minimum MAP Penalty Function** The minimum Mean Arterial Pressure (MAP) penalty function ensures MAP values remain above 60 mmHg:

$$P_{\text{minMAP}}(MAP) = \text{ReLU}\left(\frac{7(60 - MAP)}{20}\right) \tag{13}$$

  This function applies a linear penalty when MAP falls below 60 mmHg, with the penalty increasing as MAP decreases further from this threshold.

- **Pulsatility Penalty Function** The pulsatility penalty function maintains pulsatility within the range $[20, 50]$:

$$P_{\text{pulsat}}(p) = \text{ReLU}\left(\frac{7(20 - p)}{20}\right) + \text{ReLU}\left(\frac{p - 50}{20}\right) \tag{14}$$

  This bi-directional penalty function penalizes pulsatility values below 20 and above 50, with zero penalty for pulsatility in the range $[20, 50]$.

- **Hypertension Penalty Function** The hypertension penalty function penalizes elevated mean MAP values above 115 mmHg:

$$P_{\text{hyp}}(MAP) = \text{ReLU}\left(\frac{MAP - 106}{18}\right) \tag{15}$$

This function applies a linear penalty for mean MAP values exceeding the hypertension threshold of 106 mmHg.

The overall reward function combines all penalty components and negates the sum to create a reward signal:

$$R(s) = -\left[P_{\text{minMAP}}(\min(\text{MAP})) + P_{\text{hyp}}(\overline{\text{MAP}}) + P_{\text{hr}}(\min(HR)) + P_{\text{pulsat}}(\min(\text{Pulsat}))\right] \tag{16}$$

where:

- $\min(\text{MAP})$, $\min(HR)$, $\min(\text{Pulsat})$ are the minimum values over the time horizon
- $\overline{\text{MAP}}$ is the mean MAP over the time horizon
- The negative sign converts penalties into rewards (higher rewards for lower penalties)

Appendix B. Markov Decision Process (MDP) Design Details for RL

Observations. The observation space includes 12 hemodynamic features of the patient. Our inputs are the pump pressure, pump speed, and motor current 25 Hz signals recorded by the MCS device. We down-sample patient data from 25Hz to 0.00167Hz (1 sample per 10 minutes) and process them into sliding windows of 1 hour (6 time steps) to be used as states for digital twin prediction and decision making based on expert suggestion. Therefore, the observation space is $\mathcal{S} = \mathbb{R}^{6\times 12}$, where each $s_i = x_{t:t+6}$ at some t for a patient.

In – out horizon	15min – 15min 1 sample / 30s 30ts -> 30ts	1hr – 1hr 1 sample / 5min 12ts -> 12ts	1hr – 1hr 1 sample / 10min 6ts -> 6ts	2hr – 2hr 1 sample / 5min 24ts -> 24ts	2hr – 2hr 1 sample / 10min 12ts -> 12ts
MSE	0.234	0.142 ± 0.012	**0.124 ± 0.027**	0.215 ± 0.009	0.159 ± 0.006
MAP MSE	3.03	2.711 ± 0.182	**2.59 ± 0.154**	3.583 ± 0.340	3.356 ± 0.226

Table 6: Alternative settings for the digital twin. Takeaway: shorter horizon and higher down-sampling produce stronger models, but need at least 1 hour of history to provide reasonable action frequency and physiological context.

Action. The action for our MDP is the pump support level (P-level) of the MCS device. The device operates at 8 different speed levels, from P2-P9, each with a constant motor speed (rpm). The P-level proportionally determines the blood flow provided to the patient by the motor's speed and current. Clinicians can control the P-level while the patient is on support. The P-level generally stays unchanged in 1-hour intervals, unlike the state features, since it is manually controlled by the clinicians during the treatment. In practice, we take the mean P-level over the 1-hour interval as expert action. As a result, we define $\mathcal{A} = \{2, \ldots, 9\}$.

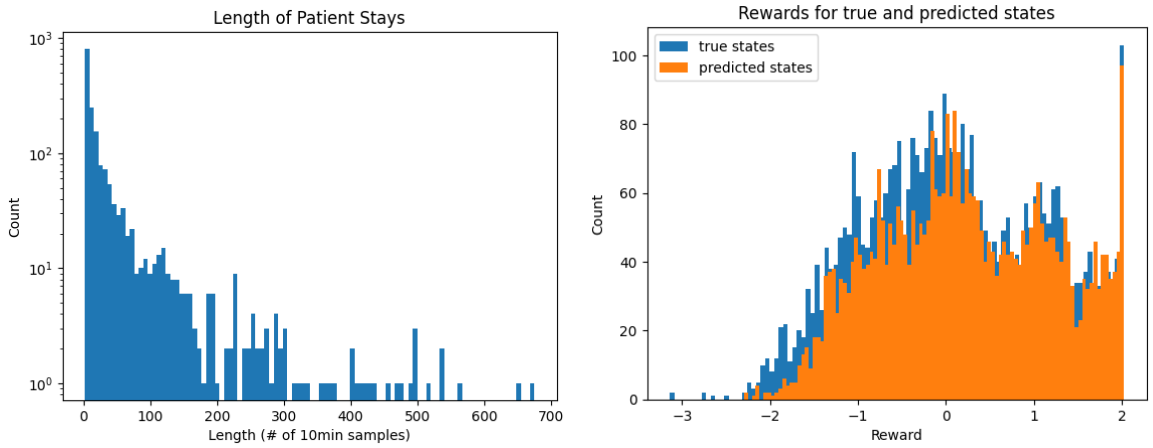


Figure 6: Length distributions of our patient data (left) and the reward score distributions of predicted states versus real patient states.

Rewards. The design table for the reward function in Appendix A is generated in line with medical consultancy. It assigns a (inverted) risk score based on acceptable intervals for hemodynamic features. The physiological reward is further normalized through Z-score normalization and clipped between $[-2, 2]$ to ensure training stability.

Challenges of Offline RL for MCS The commonly encountered issue of Offline RL is the limited access to the online environment, which results in distribution shift and large value overestimation errors to account for the shift in the real environment. While these are widely studied problems in RL, medical decision-making introduces other problems: error-prone behavioral policies, highly imbalanced actions in the dataset, and non-differentiable reward functions.

As there is no golden recipe for weaning a patient from an MCS device, the behavioral policy and the clinician policies are naturally imperfect. To this end, we expect offline RL to reveal the true policy from the hemodynamic features. Since it is required to simulate the real environment, we largely rely on a digital twin transformer model. However, the model learns to cheat by outputting cardiac cycles copied from the observation distribution. Furthermore, the action space is by definition fully constant in a state, unlike the observation space, which challenges the model compatibility.

Example weaning. We show two examples of doctors' weaning over the course of 24 hours in Figure 7.

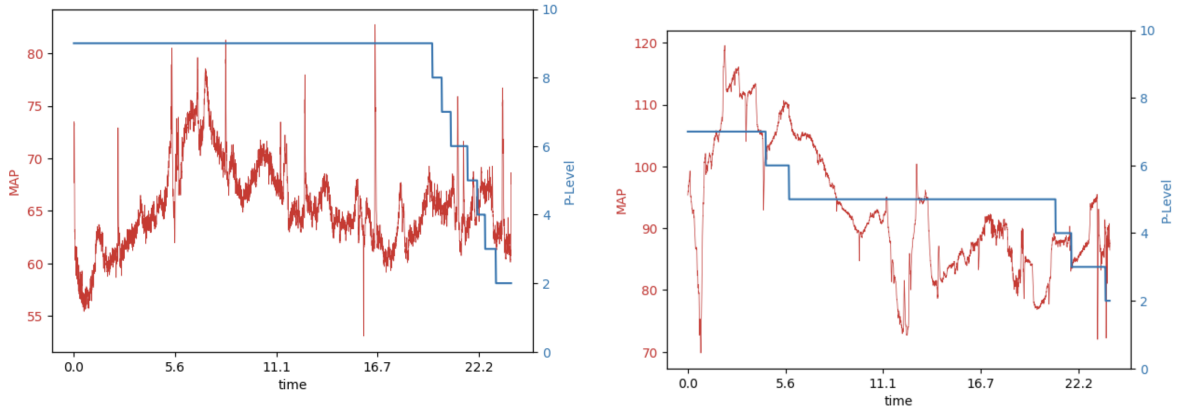


Figure 7: Example weaning of two patients over 24 24-hour horizon.

Appendix C. Algorithm Pseudocode

Algorithm 1: Clinically-aware OOD-regularized Model-based Policy Optimization (CORMPO)

Input: Dataset $D_{\text{env}} = \{(s_i, a_i, r_i, s_i')\}_i$, hyperparameters $\lambda_1, \lambda_2, \lambda$, density threshold τ
Output: Learned policy $\hat{\pi}$
`// Learn Transformer-based Digital Twin`
Train TDT $\hat{T} : S \times A \to \Delta(S)$ on D_{env} using MSE loss with dropout $p = 0.1$;
`// Learn Density Guardian`
Compute KDE density estimator: $p_{\text{KDE}}(s, a) = \frac{1}{N} \sum_{i=1}^{N} K_h((s, a) - (s_i, a_i))$;
Define low-density penalty: $u(s, a) = \tau - \log p_{\text{KDE}}(s, a)$;
`// Create Reward-Shaped Buffer`
Initialize $\mathcal{D}_{\text{shaped}} \leftarrow \emptyset$;
for *each* $(s_i, a_i, r_i, s_i') \in \mathcal{D}_{env}$ **do**
  Compute clinical metrics: $\text{ACP}(a_{i-1}, a_i)$, $\text{WS}(s_i, a_{i-1}, a_i)$;
  Shape reward: $r(s_i, a_i) \leftarrow r_{\text{phys}}(s_i) - \lambda_1 \text{ACP}(a_{i-1}, a_i) + \lambda_2 \text{WS}(s_i, a_{i-1}, a_i)$;
  Update buffer: $\mathcal{D}_{\text{shaped}} \leftarrow \mathcal{D}_{\text{shaped}} \cup \{(s_i, a_i, \tilde{r}(s_i, a_i), s_i')\}$;
end
`// Learn Policy with OOD Regularization`
Define penalized MDP $\tilde{M} = (S, A, \hat{T}, \tilde{r}, \tilde{\mu}_0, \gamma)$ with:;
$\tilde{r}(s, a) \leftarrow r(s, a) - \lambda u(s, a)$ where $\lambda = \gamma c \cdot C_{\hat{T}}$;
Train policy $\hat{\pi}$ using model-based RL (MBPO) on $\tilde{M}$;
$\hat{\pi} \leftarrow \arg\max_\pi \eta_{\tilde{M}}(\pi)$;
`// Evaluate Policy`
Evaluate $\hat{\pi}$ in TDT environment using clinical metrics (Physiological reward, ACP, WS);
return $\hat{\pi}$

Appendix D. Proofs and supporting theoretical results

Lemma 1 (Telescoping lemma - Lemma 4.1 in Yu et al. (2020)). *Let M and $\widehat{M}$ be two MDPs with the same reward function r, but different dynamics T and $\widehat{T}$ respectively. Let*

$$G^{\pi}_{\widehat{M}}(s, a) := \mathbb{E}_{s' \sim \widehat{T}(s,a)}\left[V^{\pi}_{M}(s')\right] - \mathbb{E}_{s' \sim T(s,a)}\left[V^{\pi}_{M}(s')\right].$$

Then,

$$\eta_{\widehat{M}}(\pi) - \eta_{M}(\pi) = \gamma\, \mathbb{E}_{(s,a) \sim \rho^{\pi}_{\widehat{T}}}\left[G^{\pi}_{\widehat{M}}(s, a)\right] \tag{17}$$

As an immediate corollary, we have

$$\begin{aligned} \eta_M(\pi) &= \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}\Big[r(s, a) - \gamma G^{\pi}_{\widehat{M}}(s, a)\Big] \\ &\geq \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}\Big[r(s, a) - \gamma\big|G^{\pi}_{\widehat{M}}(s, a)\big|\Big] \end{aligned} \tag{18}$$

Leveraging properties of $V^{\pi}_{\mathcal{M}}$, we will replace $G^{\pi}_{\mathcal{M}}$ by an upper bound that depends solely on the error of the dynamics T. We first note that if $\mathcal{F}$ is a set of functions mapping S to $\mathbb{R}$ that contains $V^{\pi}_{\mathcal{M}}$ then,

$$\big|G^{\pi}_{\widehat{M}}(s, a)\big| \leq \sup_{f \in \mathcal{F}} \Big|\mathbb{E}_{s' \sim \widehat{T}(s,a)}[f(s')] - \mathbb{E}_{s' \sim T(s,a)}[f(s')]\Big| =: d_{\mathcal{F}}\Big(\widehat{T}(s, a), T(s, a)\Big), \tag{19}$$

D.1. Proof of Theorem 1.

Starting from Lemma 1 (Lemma 4.1 in Yu et al. (2020), known as the telescoping lemma):

$$\eta_{\mathcal{M}}(\pi) = \eta_{\hat{\mathcal{M}}}(\pi) - \gamma \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[G^{\pi}_{\hat{\mathcal{M}}}(s,a)] \geq \eta_{\hat{\mathcal{M}}}(\pi) - \gamma \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[|G^{\pi}_{\hat{\mathcal{M}}}(s,a)|] \tag{20}$$

By Assumption 2:

$$\geq \eta_{\hat{\mathcal{M}}}(\pi) - \gamma c C_{\hat{T}} \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[u_+(s,a)] - \gamma c \epsilon_0 \tag{21}$$

Since $\eta_{\hat{\mathcal{M}}}(\pi) = \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[r(s,a)]$, add and subtract $\lambda \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[u(s,a)]$:

$$= \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[\bar{r}(s,a)] + \lambda \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[u_+(s,a) + u_-(s,a)] \\ - \gamma c C_{\hat{T}} \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[u_+(s,a)] - \gamma c \epsilon_0 \tag{22}$$

Regrouping terms and using $u_-(s,a) \leq 0$:

$$\geq \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[\bar{r}(s,a)] + (\lambda - \gamma c C_{\hat{T}}) \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[u_+(s,a)] - \lambda \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\hat{T}}}[|u_-(s,a)|] - \gamma c \epsilon_0 \tag{23}$$

Summing over the infinite horizon:

$$= \eta_{\bar{\mathcal{M}}}(\pi) - \frac{\gamma c \epsilon_{\text{approx}}}{1-\gamma} - \frac{\lambda - \gamma c C_{\hat{T}}}{1-\gamma} \mathbb{E}_{\rho^{\pi}_{\hat{T}}}[u_+] - \frac{\lambda}{1-\gamma} \mathbb{E}_{\rho^{\pi}_{\hat{T}}}[|u_-|] \tag{24}$$

Using $\lambda = \gamma c C_{\hat{T}}$,

$$= \eta_{\bar{\mathcal{M}}}(\pi) - \frac{\gamma c \epsilon_{\text{approx}}}{1-\gamma} - \frac{\lambda}{1-\gamma} \mathbb{E}_{\rho^{\pi}_{\hat{T}}}[|u_-|] \tag{25}$$

This completes the proof. □

Lemma 2 (Discounted bound for signed shaping)**.** *Starting from Assumption 1, for all* (s,a), $u(s,a) \geq u_-(s,a) \geq -\|u_-\|_\infty$. *Thus, for any policy* π, *any dynamics, and any* $\gamma \in (0,1)$,

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t u(s_t,a_t)\right] \;\geq\; -\sum_{t=0}^{\infty} \gamma^t \|u_-\|_\infty \;=\; -\frac{\|u_-\|_\infty}{1-\gamma}.$$

Consequently, for any $\lambda \geq 0$,

$$\lambda\, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t\, u(s_t,a_t)\right] \;\geq\; -\frac{\lambda}{1-\gamma}\, \|u_-\|_\infty.$$

D.2. Proof of Theorem 2.

We first note that a two-sided bound follows from Lemma 1 and Assumption 2:

$$\left|\eta_{\widehat{\mathcal{M}}}(\pi) - \eta_{\mathcal{M}}(\pi)\right| \leq \gamma c\, \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}\big[d_{\mathcal{F}}(\widehat{T}(s,a), T(s,a))\big] \leq \gamma c\, C_{\widehat{T}}\, \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[u_+(s,a)] + \frac{\gamma c}{1-\gamma} \epsilon_{\text{approx}}. \tag{26}$$

Next, recall that $\widehat{\pi}$ is optimal in the penalized MDP $\widehat{\mathcal{M}}$:

$$\widehat{\pi} = \arg\max_{\pi}\, \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}\big[r(s,a) - \lambda\, u(s,a)\big]. \tag{27}$$

Thus, for any policy π,

$$\eta_{\mathcal{M}}(\widehat{\pi}) \geq \eta_{\widehat{\mathcal{M}}}(\widehat{\pi}) - \frac{\gamma c}{1-\gamma}\epsilon_{\text{approx}} \quad \text{(by eqn 26)} \tag{28}$$

$$\geq \eta_{\widehat{\mathcal{M}}}(\pi) - \frac{\gamma c}{1-\gamma}\epsilon_{\text{approx}} \quad \text{(by optimality of } \widehat{\pi} \text{ in } \widehat{\mathcal{M}}) \tag{29}$$

$$= \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[r(s,a)] - \lambda\,\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[u(s,a)] - \frac{\gamma c}{1-\gamma}\epsilon_{\text{approx}} \quad \text{(by eqn 27)} \tag{30}$$

$$= \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[r(s,a)] - \lambda\,\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[u_+(s,a)] - \lambda\,\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[u_-(s,a)] - \frac{\gamma c}{1-\gamma}\epsilon_{\text{approx}} \tag{31}$$

$$\geq \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[r(s,a)] - \lambda\,\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[u_+(s,a)] - \frac{\gamma c}{1-\gamma}\epsilon_{\text{approx}} - \frac{\lambda}{1-\gamma}\|u_-\|_{\infty}, \tag{32}$$

where the last inequality follows from the bounded discounted sum of negative bonuses (Lemma 2), i.e., $\mathbb{E}\big[\sum_{t\geq 0}\gamma^t u_-(s_t,a_t)\big] \geq -\frac{\|u_-\|_{\infty}}{1-\gamma}$. Finally, using $\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[u_+(s,a)] \leq \delta$ for any policy π within the model-error budget δ, we obtain:

$$\eta_{\mathcal{M}}(\widehat{\pi}) \geq \max_{\pi:\,\mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[u_+]\leq\delta} \mathbb{E}_{(s,a)\sim\rho^{\pi}_{\widehat{T}}}[r(s,a)] - 2\lambda\delta - \frac{\gamma c}{1-\gamma}\epsilon_{\text{approx}} - \frac{\lambda}{1-\gamma}\|u_-\|_{\infty}, \tag{33}$$

for all $\delta \geq \delta_{\min}$, which completes the proof. □

D.3. Transformer Approximation Bound for Transition Functions

Let π be any feasible solution and $\hat{T}_\theta$ be the transformer-learned transition function with parameters θ. Assume the following:

- Dataset $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^{N}$ with N samples
- Transformer with L layers, dimension d_{model}, and H_{attn} attention heads
- Lipschitz continuous true transition T with constant L_T

Then, with probability at least $1 - 2\beta - 4\delta$, the following holds:

$$\left|V^{\pi}_{\phi,\hat{T}_\theta}(\rho_0) - V^{\pi}_{\phi,T}(\rho_0)\right| \leq \epsilon_H + \epsilon_{trans}$$

where:

$$\epsilon_H := \frac{\gamma^{H+1}(2-\gamma)\phi_{max}}{(1-\gamma)^2}$$

$$\epsilon_{trans} := \frac{\phi_{max}(\gamma - \gamma^{H+2})}{(1-\gamma)^2}\left(\epsilon_{approx} + \epsilon_{gen}\right)$$

with:

$$\epsilon_{approx} := C_{trans} \cdot \min\left\{\frac{1}{L \cdot d_{model}}, \frac{1}{H_{attn} \cdot N_{ctx}}\right\}$$

$$\epsilon_{gen} := L_T\sqrt{\frac{d\log(1/\delta) + \log N}{N}} + \mathcal{O}\left(\frac{1}{\sqrt{N_{eff}}}\right)$$

Assumption 2 is supported by the above derivation of transformer approximation error. In regions with high KDE density $p_{\text{KDE}}(s,a)$, the local sample density is high, yielding large N_{eff} and small error. Conversely,

as $p_{\text{KDE}}(s,a) \to 0$, the effective sample size diminishes, causing the error to grow. The linear relationship $\tau - p_{\text{KDE}}(s,a)$ captures this first-order dependence between local data density and model error, while $\epsilon_{\text{approx}} = C_{\text{trans}} \cdot \min\{1/(L \cdot d_{\text{model}}), 1/(H_{\text{attn}} \cdot N_{\text{ctx}})\}$ represents the transformer's irreducible approximation error determined by its architecture (depth L, dimension d_{model}, and attention heads H_{attn}). The constant $C_{\hat{T}}$ encapsulates the Lipschitz constant L_T of the true dynamics and the dimensionality-dependent factors.

Appendix E. Digital Twin Experiment Details

E.1. Digital Twin Experiments

We show experiment results of various digital twin models in Table 7. The transformer based digital twin with sinusoidal positional embeddings outperforms all baselines, including the transformer architecture employing rotary positional embeddings. Figure 8 provides an analysis of the error accumulation over prediction horizon for our digital twin - for the three metrics that we base our reward on, the error accumulates sub-linearly and is low even at the 6 hour horizon.

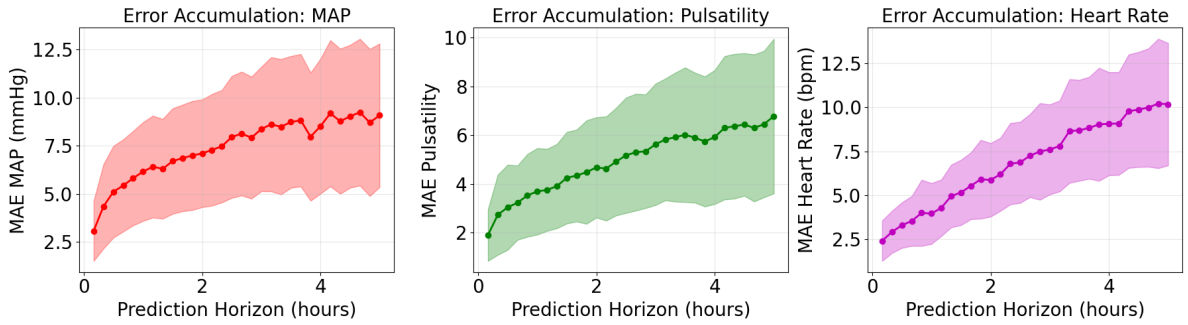


Figure 8: Error accumulation of the digital twin predictions (mean ± standard deviation evaluated across 500 samples). We can see that the error, especially mean average error of MAP, accumulates slowly through the horizon.

	MAE	MAE (MAP only)	MAE Static PL	MAE changing PL	Trend Acc.	CRPS
MLP	9.85 ± 0.44	4.11 ± 0.01	8.88 ± 0.45	13.76 ± 0.40	0.83 ± 0.03	7.43 ± 0.22
Neural Process	8.32 ± 0.18	4.63 ± 0.06	6.83 ± 0.26	14.31 ± 0.22	0.89 ± 0.00	4.92 ± 0.66
CLMU	7.61 ± 0.12	4.31 ± 0.04	7.00 ± 0.11	10.06 ± 0.17	0.89 ± 0.00	5.48 ± 0.09
SSM	8.12 ± 0.46	4.12 ± 0.11	7.49 ± 0.55	10.65 ± 0.11	0.88 ± 0.00	4.43 ± 0.29
TDT (ours)	**5.41 ± 0.05**	**3.88 ± 0.12**	**4.90 ± 0.05**	**7.47 ± 0.08**	0.88 ± 0.01	**3.45 ± 0.12**

Table 7: Digital twin model evaluation; Transformer outperforms baselines in all metrics.

E.2. Baselines

We evaluate our approach against several established baselines for probabilistic dynamics modeling. Each baseline is configured with carefully tuned hyperparameters to ensure fair comparison:

- **Multi-Layer Perceptrons (MLPs)** with Monte Carlo dropout approximate probabilistic forecasts by treating dropout as a Bayesian approximation technique, enabling uncertainty quantification through multiple forward passes during inference. The MLP baseline employs a three-layer architecture with hidden dimensions $[512, 256, 128]$, ReLU activation functions, and a dropout rate of 0.2 applied after

each hidden layer. The network flattens the input sequence and concatenates it with p-level control signals before processing through the fully connected layers.

- **Neural Processes** Sun et al. (2024) is a meta-learning approach that conditions on context observations to predict distributions over functions, enabling few-shot adaptation to new dynamical systems while maintaining uncertainty quantification. The implementation features a latent dimension of 128, a hidden dimension of 256, and employs separate encoder networks for context processing with three-layer architectures. The context encoder processes input features augmented with time indices, while the aggregator combines encoded representations across time steps. The decoder network generates both mean and variance predictions for each feature at each forecast timestep.
- **Conditional Legendre Memory Units (CLMUs)** Li et al. (2022); Voelker et al. (2019) leverage orthogonal polynomial basis functions to capture long-term temporal dependencies through structured memory mechanisms. The CLMU baseline utilizes 2 layers with memory dimension 64, hidden dimension 128, and incorporates p-level conditioning through a dedicated projection layer. Each LMU layer employs Legendre polynomial transition matrices with scaling parameter $\theta = 1.0$ and applies exponential smoothing with decay rate 0.9 for stable memory updates. The output projection includes dropout with a rate 0.1 for regularization.
- **State Space Models (SSMs)** Hamilton (1994) represent dynamics through latent state evolution governed by linear or nonlinear transition functions, naturally incorporating temporal dependencies and enabling principled probabilistic inference over hidden states. The SSM baseline operates with state dimension 64, hidden dimension 128, and forecast horizon of 6 steps. The state transition matrix is initialized as $0.9 \cdot I + 0.1 \cdot \mathcal{N}(0, 1)$ to ensure stability, while the observation model employs a two-layer network with ReLU activation and dropout rate 0.1. Stochastic sampling is achieved by injecting Gaussian noise with standard deviation 0.01 during state transitions.
- **Transformers with sinusoidal positional embeddings (TDT sin.)** Vaswani et al. (2017) and **Transformers with rotary positional embeddings (TDT rot.)** Su et al. (2024). TDT (rot.) leverages self-attention mechanisms enhanced with rotary position encoding (RoPE) that captures relative positional relationships through multiplicative rotations. The transformer model's attention mechanism allows the model to attend to relevant temporal patterns and input control, improving the time series prediction. Both transformer models have the architecture outlined in our methodology section.

All baselines are trained using the Adam optimizer with a learning rate 0.001 and employ Monte Carlo sampling with 50 forward passes for uncertainty quantification during inference.

E.3. Metrics for evaluating digital twin

- MAE All: Mean Absolute Error across all features
- MAE MAP: Mean Absolute Error for MAP (Mean Arterial Pressure) only
- MAE Static: MAE for samples with non-changing P-Levels over the course of 2 hours.
- MAE Dynamic: MAE for samples with dynamic p-levels
- Trend Acc: Trend direction accuracy for MAP. Trend is classified as (1) increasing if the slope of MAP over the predicted horizon is ≥ 2, (2) decreasing if the slope ≤ 2, and (3) flat otherwise.
- CRPS: Continuous Ranked Probability Score is a proper scoring rule Gneiting and Raftery (2007) for uncertainty quantification, calculated from 50 samples from the probabilistic predictions (samples x, x') and the ground truth y as in equation 34.

$$\mathrm{CRPS}(\mathcal{F}, y) = \int \left(\mathcal{F}(x) - \mathbf{1}\{x \geq y\}\right)^2 dx = \mathbb{E}_x[|x - y|] - \frac{1}{2}\mathbb{E}_{x,x'}[|x - x'|] \tag{34}$$

E.4. Additional Visualizations.

Please see figure 9 for further qualitative examples of digital twin models.

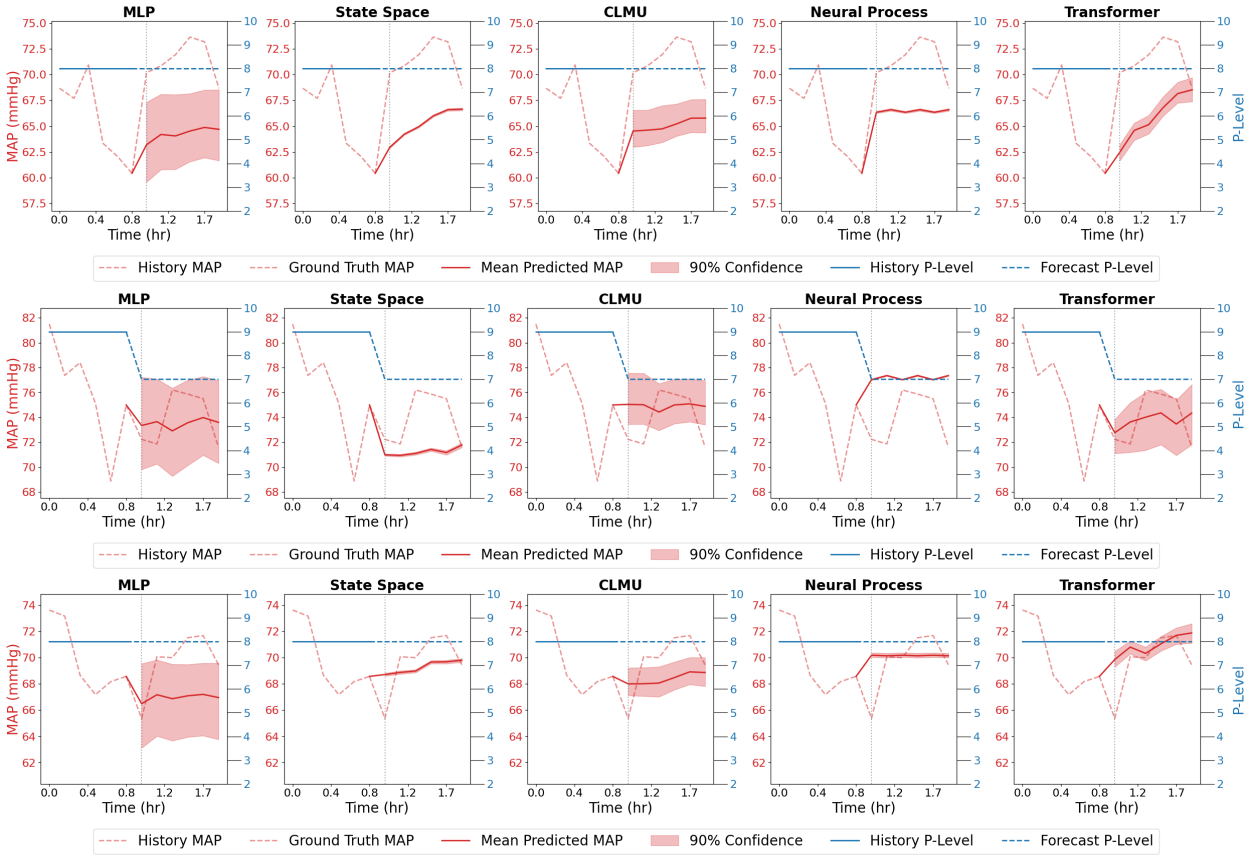


Figure 9: Digital twin prediction visualization compared with baselines. The Transformer model is more accurate in reflecting response to P-level change and more expressive when capturing large changes in patient state, resulting in its higher accuracy.

Appendix F. RL Experiment Settings and Additional Results

F.1. Implementation details and hyperparameters of CORMPO

We first start our implementation by generating the replay buffer rewards shaped with λ_1 (ACP weight), and λ_2 (WS weight) parameters. ACP and WS takes the current state, s_t, as the stability condition and evaluates the change in action from current action, a_t, to the next action, a_{t+1}. In parallel, we apply KDE on a randomly shuffled training split comprising 80% of the full dataset and evaluate thresholds on a held-out 10% validation set. Only for real dataset, we choose the threshold percentile as 35% due to the distribution discrepancy between train and validation. For synthetic dataset, the optimal threshold is selected by searching across percentiles to on the validation set by minimizing the log-likelihood variance on the ID-labeled region resulting in 20% for both. For each query, we retrieve nearest neighbors from the training data and fit a kernel to compute log-density scores. Merging the two modules, we deploy KDE model and the selected threshold as the penalization on the rewards predicted by the dynamics model during MBPO training with the reward-shaped replay buffer.

Parameters	Value
Actor learning rate	3×10^{-4}
Critic learning rate	3×10^{-4}
Discount factor (γ)	0.99
Target network update coefficient (τ)	0.005
Target entropy (often $-$action dimension)	-1
Temperature optimizer learning rate	3×10^{-4}
Dynamics model learning rate	1×10^{-3}
Dynamics ensemble size	7
Holdout ratio	0.2
Training epochs	100
Steps per epoch	1000
Evaluation episodes	1000
Mini-batch size	256
Model rollout horizon	5
Rollout batch size	10000
Rollout frequency	1000
Real-to-model data sampling ratio	0.05

Table 8: Base hyperparameters of our CORMPO implementation.

KDE Hyperparameters. We selected the RBF kernel for this method. The bandwidth and number of neighbors are selected based on the modeled distribution on the validation set. As we increase the bandwidth, the distribution starts to be sharper. So, we chose 1 as the bandwidth for KDE. We select 100 neighbors, as our training set includes 12051 samples.

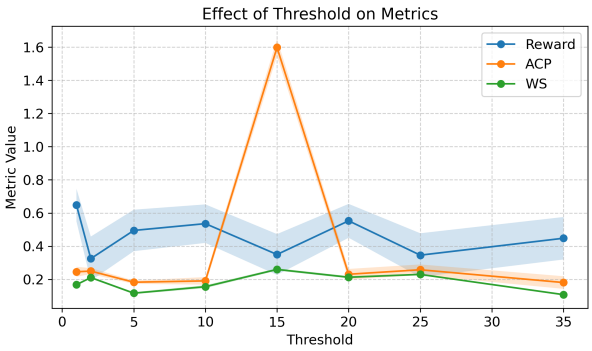
CORMPO Hyperparameters. Our implementation of model-based policy optimization is largely borrowed by Yu et al. (2020) (see Table 8) with the change of reward penalization in the dynamics model roll-outs. We select the hyperparameters of λ_1, λ_2, and λ by evaluating each model in the TDT environment. Since there is no golden aggregation of the medical metrics, we pick the model with the best reward, with ACP and WS outperforming the baselines.

- **Noiseless Dataset Experiment**: $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, $\lambda = 0.005$, epoch $= 90$.
- **Noisy Dataset Experiment**: $\lambda_1 = 0.0$, $\lambda_2 = 0.0$, $\lambda = 0.08$, epoch $= 100$.
- **Real-life Dataset Experiment**:$\lambda_1 = 1.0$, $\lambda_2 = 0.0$, $\lambda = 0.005$, epoch $= 100$.

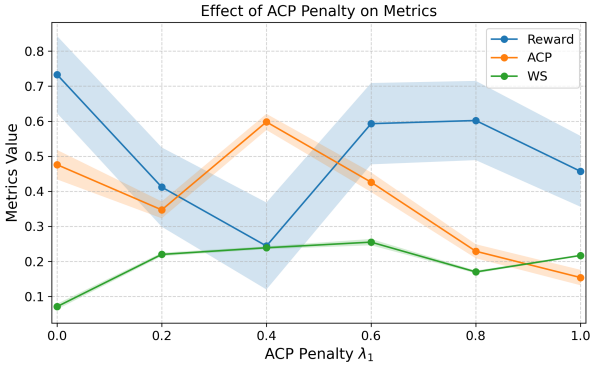
F.2. Additional Results

Penalty type	Expert DT	BC	MBPO	MOPO	SVR	CORMPO *KDE*	CORMPO *RealNVP*
Reward	0.557	0.175 ± 0.118	0.420 ± 0.139	0.373 ± 0.129	0.530 ± 0.152	$\mathbf{0.687 \pm 0.106}$	0.516 ± 0.126
ACP	1.79	0.068 ± 0.012	0.459 ± 0.032	0.984 ± 0.020	0.599 ± 0.057	$\mathbf{0.018 \pm 0.007}$	0.258 ± 0.036
WS	0.053	$\mathbf{0.345 \pm 0.008}$	0.147 ± 0.007	0.042 ± 0.006	0.166 ± 0.003	0.173 ± 0.007	0.155 ± 0.003

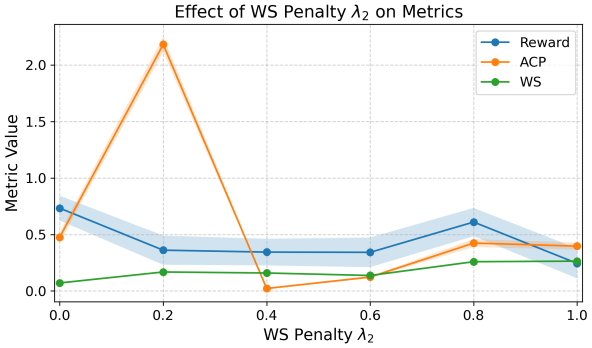
Table 9: Comparison of CORMPO based on KDE, and CORMPO based on RealNVP against baseline models with 1000 episodes averaged over 5 seeds. Evaluation is completed in the noiseless environment setting. CORMPO with RealNVP does not outperform the baselines, showing moderate performance in all metrics. This result suggests further investigations on calibrating CORMPO to different density estimators.



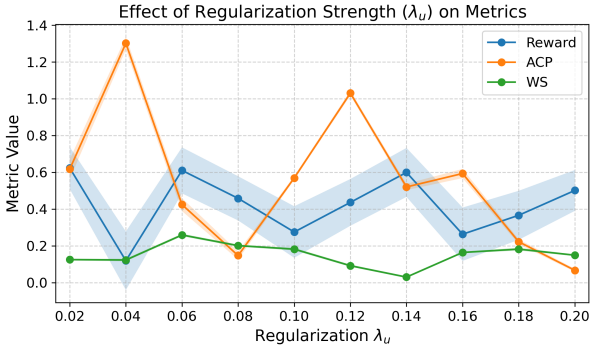
(a) Effect of threshold τ. The x-axis is the % of validation data flagged anomalous.



(b) Effect of ACP penalty λ_1 ($\lambda_2 = 0$, $\tau = 0.005$).



(c) Effect of WS penalty λ_2 ($\lambda_1 = 0$, $\tau = 0.005$).



(d) Effect of regularizer λ_u ($\lambda_1 = 0$, $\lambda_2 = 0.8$).

Figure 10: Sensitivity analysis of CORMPO under (a) density thresholds, (b–c) penalty coefficients, and (d) regularization.

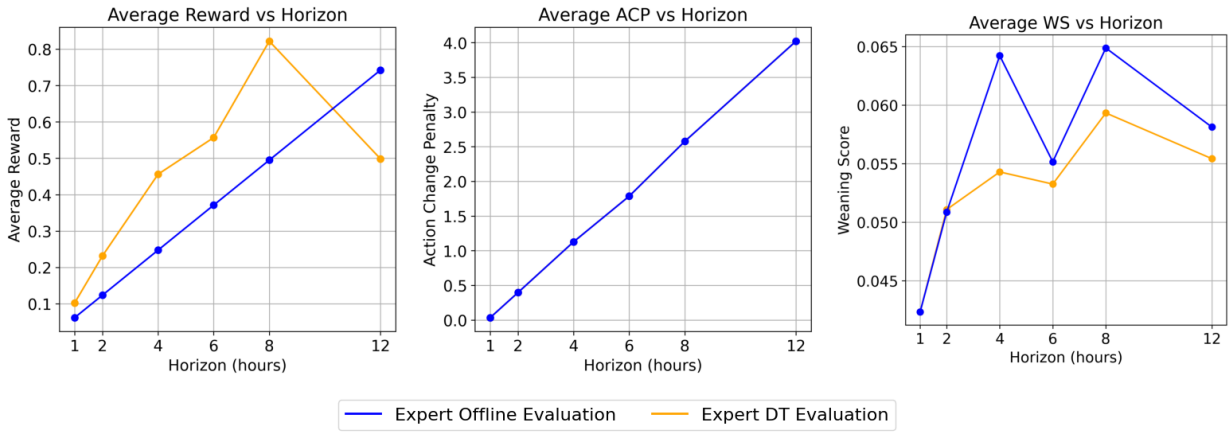


Figure 11: Comparison of change in physiological reward, ACP, and WS metrics with respect to various evaluation horizons in hours. This experiment is done on the expert (clinician) decisions evaluated by our digital twin (orange) and directly on the offline patient outcomes (blue). We observe that our digital twin performs similar to the offline dataset. Noting that digital twin results in mostly higher physiological rewards, the gap between offline and digital twin evaluation increases within a tolerable bound, where the largest increase is 60%. ACP remains the same since the compared actions are the real p-levels while only the next state differs. In WS, the digital twin depicts a small decrease, which is always bounded until the 12-hour horizon.

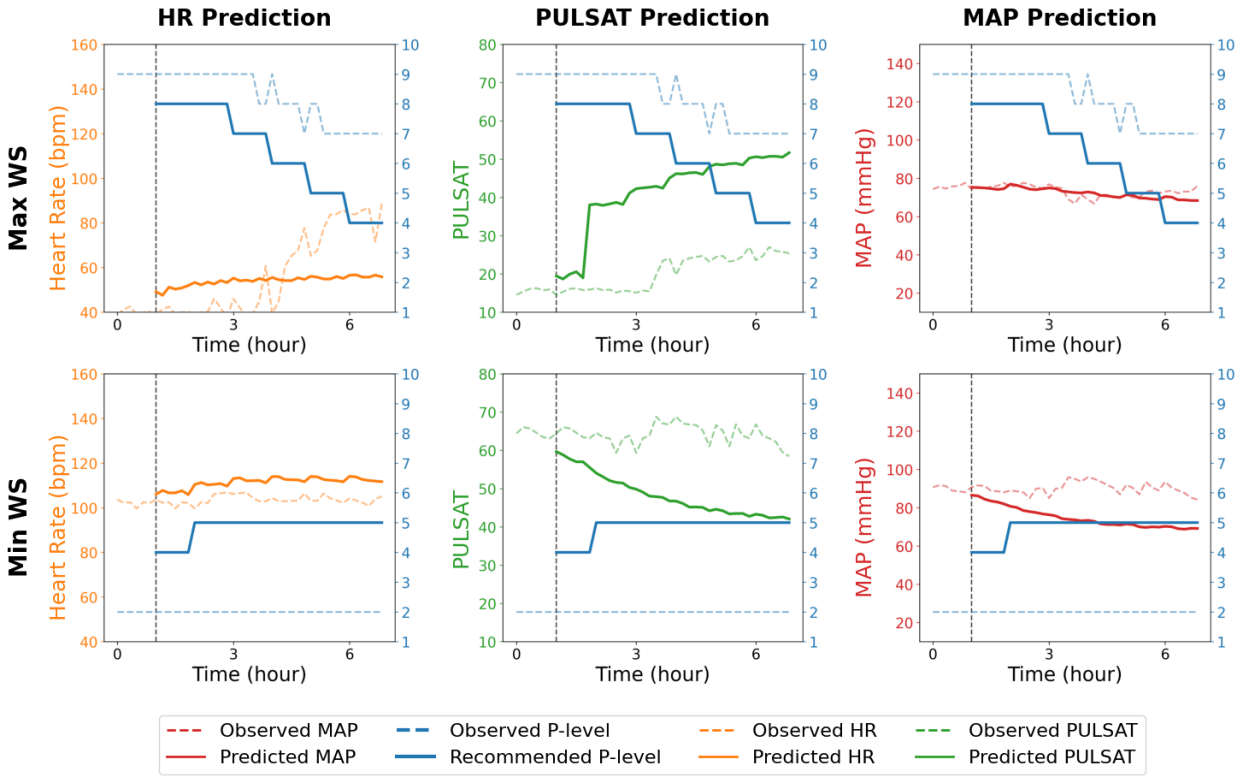


Figure 12: Trained on the real-life dataset, we compare our CORMPO against baselines in 6-hour digital twin rollouts. The upper row depicts high (0.75) and the lower row depicts low (-0.334) WS roll-outs with p-level recommendations on the real data. In the highest WS row, we predict all vitals to remain within safe hemodynamic regions (see Table 5), with HR and MAP showing stationary behavior and pulsatility rising above the critical threshold. We observe decaying pulsatility and MAP in the low WS case.

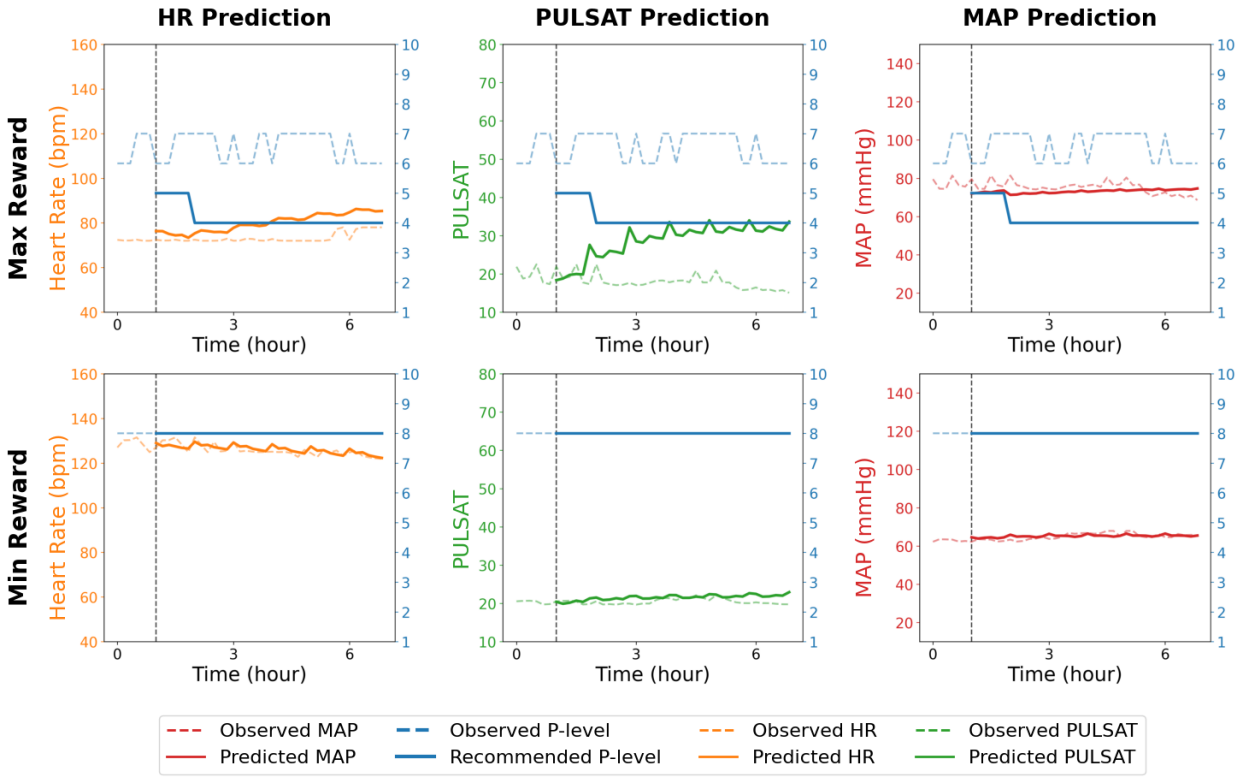


Figure 13: Trained on the real-life dataset, we compare our CORMPO against baselines in 6-hour digital twin rollouts. The upper row corresponds to the high physiological reward roll-outs (3.84), and the lower row to the low physiological reward roll-outs (–12.0). In the lower row, pulsatility and MAP are on the critical threshold. In the upper row, heart rate and pulsatility increase and stabilize away from the critical thresholds. As MAP is also predicted to remain stable, CORMPO initiates the weaning process.

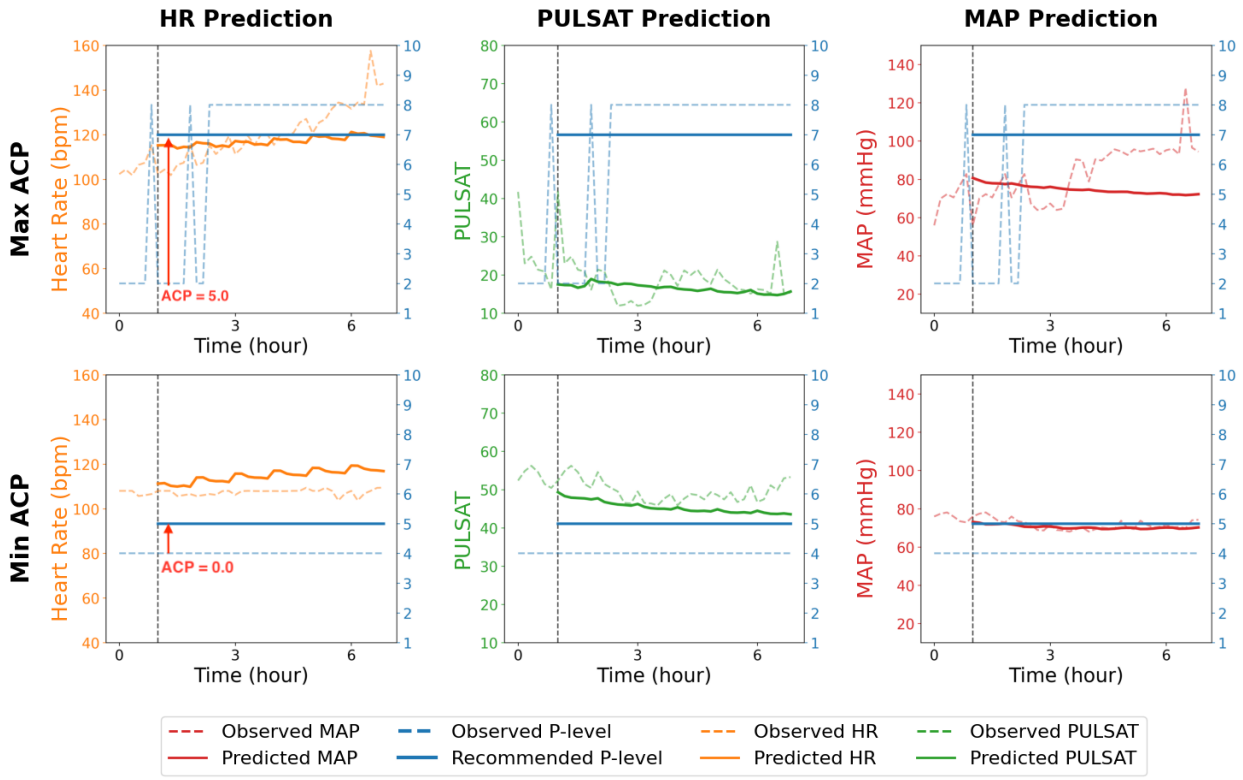


Figure 14: Trained on the real-life dataset, we compare our CORMPO against baselines in 6-hour digital twin rollouts. We depict high (5.00) and low (0.00) ACP roll-outs with p-level recommendations. To note, our ACP definition states $\text{ACP}_t = |a_{t+1} - a_t|$ only if $\text{ACP}_t > 2$ where a_0 is the action observed at time $t = 1$ in the plots. Naturally, the upper row results in 5.0 ($|a_1 - a_0| = 5$) while the last row results in 0 ACP as indicated with the red arrow. In the lower row, with the p-level increased by 1 relative to the ground truth, the vitals remain stationary and away from the critical thresholds. In the upper row, ACP is 5.0 since the policy increases the p-level by 5 relative to the ground truth p-level. In this case, pulsatility and MAP decay down to the critical threshold.