Treatment Stitching with Schrödinger Bridge for Enhancing Offline Reinforcement Learning in Adaptive Treatment Strategies

Dong-Hee Shin, Deok-Joong Lee, Young-Han Son, Tae-Eui Kam

Department of Artificial Intelligence, Korea University
Seoul, Republic of Korea
{dongheeshin, deokjoong, yhson135, kamte}@korea.ac.kr

Abstract

Adaptive treatment strategies (ATS) are sequential decision-making processes that enable personalized care by dynamically adjusting treatment decisions in response to evolving patient symptoms. While reinforcement learning (RL) offers a promising approach for optimizing ATS, its conventional online trial-and-error learning mechanism is not permissible in clinical settings due to risks of harm to patients. Offline RL tackles this limitation by learning policies exclusively from historical treatment data, but its performance is often constrained by data scarcity—a pervasive challenge in clinical domains. To overcome this, we propose *Treatment Stitching* (*TreatStitch*), a novel data augmentation framework that generates clinically valid treatment trajectories by intelligently stitching segments from existing treatment data. Specifically, *TreatStitch* identifies similar intermediate patient states across different trajectories and stitches their respective segments. Even when intermediate states are too dissimilar to stitch directly, *TreatStitch* leverages the Schrödinger bridge method to generate smooth and energy-efficient bridging trajectories that connect dissimilar states. By augmenting these synthetic trajectories into the original dataset, offline RL can learn from a more diverse dataset, thereby improving its ability to optimize ATS. Extensive experiments across multiple treatment datasets demonstrate the effectiveness of *TreatStitch* in enhancing offline RL performance. Furthermore, we provide a theoretical justification showing that *TreatStitch* maintains clinical validity by avoiding out-of-distribution transitions.

Introduction

Imagine a scenario where a patient needs to visit the hospital regularly to manage chronic health diseases. During each visit, clinicians assess the patient's symptoms, review their medical history, and prescribe a treatment tailored to their current condition. After observing how the patient responds, clinicians utilize this new information to adapt treatment strategies to optimize patient outcomes. This dynamic process—where each decision is informed by a sequence of past observations, treatments, and responses—is referred to as an adaptive treatment strategy (ATS) (Murphy 2005). As depicted in Figure 1(a), ATS play a crucial role in delivering personalized care in longitudinal clinical settings, where treatment decisions adapt to a patient's evolving symptoms.

In recent years, advancements in artificial intelligence (AI) have sparked a growing interest in developing AI-driven clinical decision support systems to enhance the implementation and optimization of ATS (Giordano et al. 2021). Early research in this domain has predominantly focused on behavior cloning (BC) (Torabi, Warnell, and Stone 2018). To be specific, BC learns a policy through supervised learning, where the objective is to replicate clinicians' decisions at each time step by minimizing the discrepancy between AI-generated treatment recommendations and clinicians' actual decisions (Wang et al. 2020).

While BC is straightforward and intuitive, it prioritizes short-term alignment with clinicians' decisions rather than optimizing long-term clinical outcomes (Wang et al. 2018). This limitation implies that BC can only reproduce past clinicians' decisions and may overlook potentially superior treatment strategies that could yield better long-term outcomes. To overcome this limitation, the reinforcement learning (RL) approach has been introduced as RL can optimize for long-term outcomes by maximizing cumulative rewards (Liu et al. 2017; Zhang and Bareinboim 2019). Moreover, RL enables an agent to actively explore its environment, allowing it to discover better solutions (Shin et al. 2024a).

However, a fundamental challenge in applying RL to clinical settings arises from its learning mechanism. As shown in Figure 1(b), conventional online RL learns a policy through an online trial-and-error process. In clinical settings, this would necessitate experimenting with different treatment strategies directly on patients to evaluate their efficacy, which raises significant ethical and safety concerns (Yu et al. 2021). To address this challenge, offline RL has emerged as a promising alternative (Fatemi et al. 2022; Levine et al. 2020). As shown in Figure 1(c), offline RL utilizes extensive clinical databases to construct a static offline dataset that contains historical treatment data collected from real-world clinical practices. This offline dataset is typically composed of sequences of states (symptoms), actions (prescribed treatments), and rewards (responses), providing a rich source of information for policy learning. Next, offline RL analyzes these historical data to identify treatment decisions that maximize cumulative rewards. As a result, offline RL can provide effective treatment strategies for clinicians without the risks of online trial-and-error experimentation on patients (Luo et al. 2024; Kondrup et al. 2023).

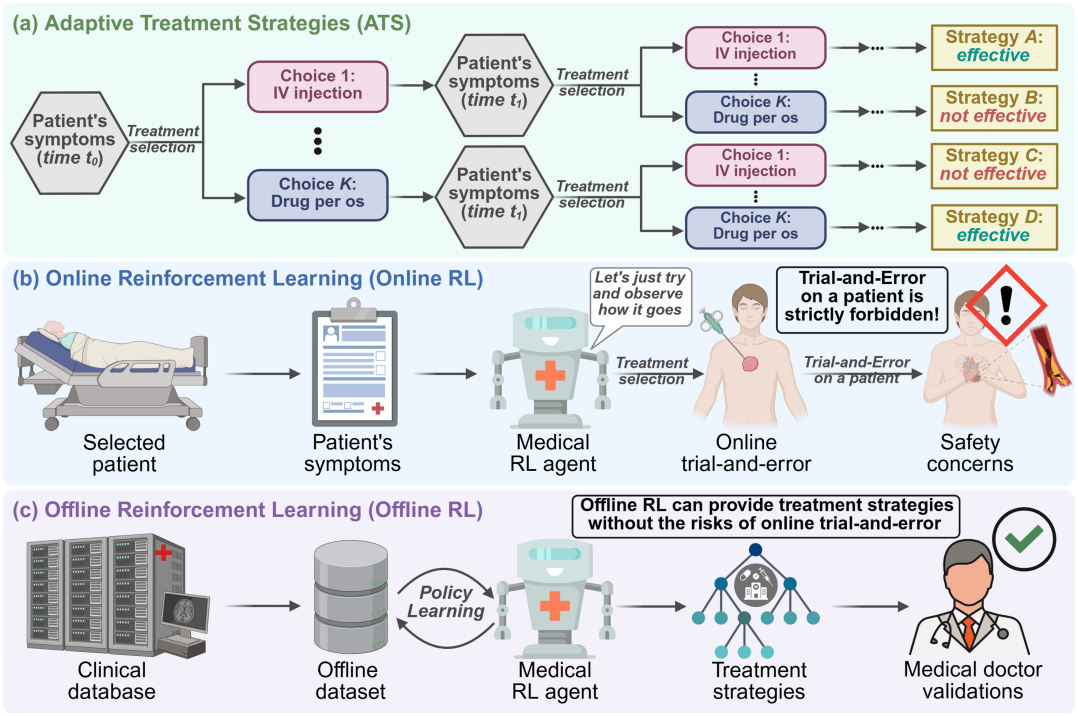

Figure 1: Illustrations of adaptive treatment strategies (ATS) and reinforcement learning (RL). (a) ATS aims to identify effective treatment strategies based on a patient's evolving symptoms. (b) Online RL learns via trial-and-error, raising safety concerns in clinical settings. (c) Offline RL learns a policy from the offline dataset, removing the need for online trial-and-error on patients.

Despite the promise of offline RL, several challenges persist in its clinical applications. A well-known limitation of RL algorithms is their substantial data requirements for effective training (Sutton and Barto 2018). This limitation implies a need for a large and comprehensive offline dataset to train robust offline RL policies. However, clinical settings are inherently data-hungry, often limited by the availability and diversity of historical treatment data (Shaheen 2021). A common approach to address this data scarcity issue is data augmentation, particularly through the use of generative models, showing promising results in domains like medical imaging (Shorten and Khoshgoftaar 2019). However, applying generative models to generate synthetic treatment data entirely from scratch presents unique challenges. Unlike medical images, treatment data often exhibit long-term causal dependencies across multiple treatment stages. Thus, generative models may struggle to capture these longitudinal dynamics, potentially leading to low-quality synthetic data. Moreover, generating synthetic data entirely from scratch can lead to error accumulation over time (Bauer et al. 2024).

In this paper, we introduce *Treatment Stitching* (*TreatStitch*), a novel data augmentation framework designed for offline RL in ATS applications. Unlike methods that rely on generative models to synthesize treatment data from scratch, *TreatStitch* generates synthetic trajectories by intelligently stitching together segments from real patient trajectories in the offline dataset. Specifically, it identifies similar intermediate patient states (e.g., similar clinical conditions) across different trajectories. When such similar states are found, *TreatStitch* 'cuts' both trajectories at that point and 'stitches' the first segment of one trajectory with the second segment of the other, creating a new, clinically valid stitched trajectory that preserves authentic state-action transitions.

However, when all trajectories do not share similar intermediate states, *TreatStitch* leverages the Schrödinger bridge method (Wang et al. 2021) to construct smooth and energy-efficient bridging trajectories between dissimilar states. This approach greatly expands data augmentation opportunities, particularly within sparse or heterogeneous offline datasets. Moreover, by restricting synthetic data generation to only these minimal bridging trajectories, we reduce the potential for error accumulation that can arise from generating extensive synthetic data from scratch. As a result, *TreatStitch* significantly enhances the diversity and coverage of the offline dataset by augmenting these clinically valid stitched trajectories. The main contributions of this work are outlined as:

- To the best of our knowledge, *TreatStitch* is the first data augmentation framework for ATS applications to utilize a trajectory stitching method that generates clinically valid synthetic trajectories from existing offline treatment data.
- Even when offline data is sparse or heterogeneous, making direct stitching between states difficult, we introduce the Schrödinger bridge to construct smooth bridging trajectories that enable stitching between dissimilar states.
- We empirically demonstrate the effectiveness of *TreatStitch* in enhancing offline RL performance and theoretically explain its capability to preserve clinical validity.

Related Work and Preliminary

Adaptive Treatment Strategies (ATS). ATS refer to clinical sequential decision-making systems that dynamically adapt treatment recommendations in response to evolving patient symptoms. Actions (i.e., treatment decisions) in ATS should take into account not only current symptoms, but also medical history and prior treatment responses, with the goal of optimizing long-term clinical outcomes rather than merely addressing immediate symptoms. In recent years, a diverse array of studies have been proposed to optimize ATS. Early methods utilized either online RL (Liu et al. 2017) or imitation learning (Wang et al. 2020) methods to tackle ATS optimization. However, these approaches have some limitations: online RL poses safety concerns due to online patient experimentation, while imitation learning merely replicates clinicians' past decisions without optimizing for long-term outcomes. To address these challenges, offline RL (Kondrup et al. 2023; Cai et al. 2023; Eghbali, Alhanai, and Ghassemi 2025) has emerged as better and safer alternative method as it enables learning optimal policies from historical treatment data without requiring online experimentation on patients. However, existing offline RL methods primarily focus on algorithmic improvements while overlooking the critical data perspective—despite the fact that RL requires large amounts of diverse data for effective training (Levine et al. 2020). Therefore, in this study, we introduce a data augmentation framework that generates clinically valid synthetic treatment data to enhance offline RL performance in ATS applications. Extended related work section is provided in Appendix A.

Markov Decision Process (MDP) Formulation in ATS. RL problems are typically formulated as an MDP, which provides a mathematical framework for modeling the decision-making process. Formally, an MDP is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{F}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space that represents the set of all possible patient symptoms; $\mathcal{A}$ is the action space, corresponding to all possible treatment decisions that clinicians can prescribe; $\mathcal{F} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function that defines the probability of transitioning to next state; $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function that maps state-action pairs to scalar rewards, often representing patient responses or overall clinical outcomes; $\gamma \in [0, 1]$ is the discount factor that determines the present value of future rewards. At each time step t, the RL agent receives the current state $s_t \in \mathcal{S}$ and executes an action $a_t \in \mathcal{A}$ according to a policy $\pi : \mathcal{S} \to \mathcal{A}$. This action leads to a next state s_{t+1} and yields a reward $r_t = \mathcal{R}(s_t, a_t)$ that quantifies the clinical outcome. The main goal of the RL agent is to find an *optimal policy* π^* that maximizes the expected discounted cumulative reward such as follows:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T} \gamma^t \mathcal{R}(s_t, a_t)\right], \tag{1}$$

where $\tau = \{(s_0, a_0, r_0), \ldots, (s_T, a_T, r_T)\}$ denotes a treatment trajectory of length T. The typical online RL allows the agent to explore its environment to gather new trajectories and update the policy. However, in clinical settings, online exploration (i.e., trial-and-error on real patients) is strictly prohibited. This constraint motivates the use of offline RL.

Offline Reinforcement Learning (RL) in ATS. In offline RL, the agent is prohibited from further interaction with the environment. Instead, the agent must learn from a static offline dataset $\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$ of N treatment trajectories collected from the historical clinical database. Thus, the main goal of offline RL is to effectively leverage the offline dataset to learn an optimal (or near-optimal) policy without further exploration. However, this offline setting introduces two key challenges: **(1)** *distributional shift*, which occurs when the learned policy suffers from unreliable evaluation due to distribution mismatch; and **(2)** *extrapolation error*, which results in overly optimistic predictions when dealing with out-of-distribution (OOD) state-action pairs (Levine et al. 2020). To address these challenges, the Conservative Q-Learning (CQL) algorithm (Kumar et al. 2020) has been proposed. Specifically, CQL introduces a conservative regularization that explicitly penalizes overly optimistic predictions by discouraging high Q-values on OOD (unseen) state-action pairs. The loss function for CQL can be expressed as:

$$\begin{aligned}\mathcal{L}(\theta) = &\underbrace{\mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\left(Q_\theta(s,a) - \hat{\mathcal{B}}Q_\theta(s,a)\right)^2\right]}_{\text{Bellman Error}} \\ &+ \beta\underbrace{\left(\mathbb{E}_{a\sim\mu}\left[Q_\theta(s,a)\right] - \mathbb{E}_{a\sim\mathcal{D}}\left[Q_\theta(s,a)\right]\right)}_{\text{Conservative Regularization}},\end{aligned} \tag{2}$$

where $Q_\theta(s, a)$ represents the current Q-value estimate and $\hat{\mathcal{B}}$ is a Bellman backup operator (Sutton and Barto 2018) that refines this Q-value estimate by considering both the immediate reward and the discounted future reward. The first Bellman error term encourages Q_θ to accurately fit the observed trajectories in the offline dataset $\mathcal{D}$. The second regularization term penalizes Q_θ for OOD actions sampled from μ, mitigating overly optimistic predictions. Note that μ typically represents the distribution for modeling OOD actions, such as a uniform distribution or the current policy π.

Method

Treatment Stitching Framework for Offline RL

To address the data scarcity challenge in clinical settings, we propose the *Treatment Stitching* (*TreatStitch*) framework, which fully leverages existing treatment trajectories in offline datasets to generate new stitched treatment trajectories. As depicted in Figure 2(a), the overall process begins with the offline dataset $\mathcal{D} = \{\tau_1, \tau_2, \ldots, \tau_N\}$, where each trajectory $\tau_i = \{(s_t^i, a_t^i, r_t^i)\}_{t=0}^{T}$ consists of a sequence of states, actions, and rewards observed in clinical database. Then, the trajectories are categorized into two groups: the high reward group $\mathcal{D}_{\text{high}}$, consisting of trajectories that lead to beneficial clinical effects with high cumulative rewards, and the low reward group $\mathcal{D}_{\text{low}}$, which include trajectories with adverse effects and low cumulative rewards. Formally, it is given by:

$$\mathcal{D}_{\text{high}} = \left\{\tau_i \in \mathcal{D} \mid \sum_{t=0}^{T_i} \gamma^t r_t^i \geq \Phi_q(\mathcal{D})\right\}, \mathcal{D}_{\text{low}} = \mathcal{D} \setminus \mathcal{D}_{\text{high}},$$

where $\Phi_q(\mathcal{D})$ denotes the reward value at the q-th percentile of cumulative rewards across all trajectories in $\mathcal{D}$.

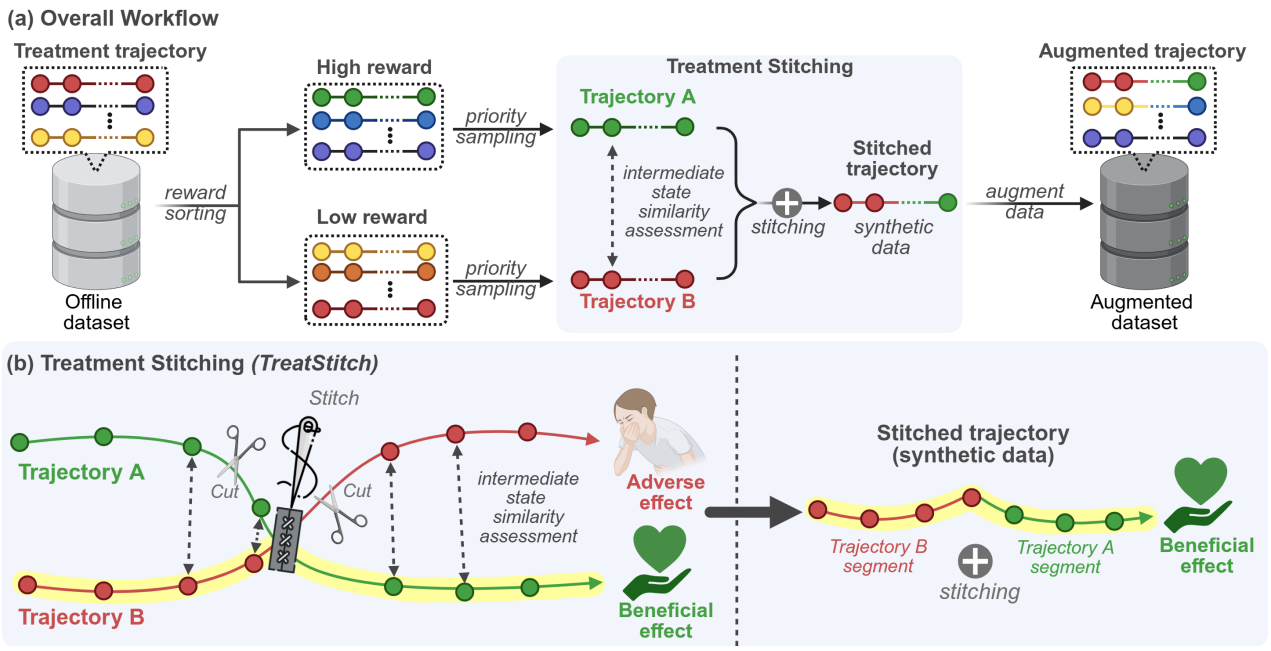

Figure 2: (a) The overall workflow of our treatment stitching framework that produces an augmented dataset for enhancing offline RL. (b) The detailed process of treatment stitching, which generates stitched trajectories from existing data.

Priority Sampling. Subsequently, we introduce priority sampling using the Boltzmann distribution to strategically select trajectories for the stitching process. The key insight is that combining trajectories with lower rewards (adverse effects) with those having higher rewards (beneficial effects) is more likely to produce informative stitched trajectories, where treatment trajectories beginning with adverse effects can still evolve into beneficial outcomes. To implement this intuition, our sampling strategy prioritizes trajectories with higher rewards in $\mathcal{D}_{\text{high}}$ and lower rewards in $\mathcal{D}_{\text{low}}$. Formally, the sampling probability of trajectory τ_i in $\mathcal{D}_{\text{high}}$ is given by:

$$p(\tau_i \mid \tau_i \in \mathcal{D}_{\text{high}}) = \frac{\exp(R(\tau_i)/\alpha)}{\sum_{\tau_j \in \mathcal{D}_{\text{high}}} \exp(R(\tau_j)/\alpha)}, \quad (3)$$

where $R(\tau_i) = \sum_{t=0}^{T} \gamma^t r_t^i$ is the cumulative reward of τ_i, and α is the temperature parameter. For trajectories in $\mathcal{D}_{\text{low}}$, we use $-R(\tau_i)$ to prioritize trajectories with lower rewards. As training progresses, we gradually increase α to shift from highly prioritized to uniform sampling for broader coverage. Further analysis of priority sampling is in Appendix D.

Treatment Stitching. Trajectory A (τ_A) and Trajectory B (τ_B) are sampled from $\mathcal{D}_{\text{high}}$ and $\mathcal{D}_{\text{low}}$, respectively, using priority sampling. To identify a potential stitching point, we perform an *intermediate state similarity assessment* by computing the cosine similarity between all pairs of intermediate states—comparing each state $\{s_t^A\}$ in τ_A with $\{s_{t'}^B\}$ in τ_B:

$$\text{Sim}\left(s_t^A, s_{t'}^B\right) = \frac{\langle s_t^A, s_{t'}^B \rangle}{\|s_t^A\| \|s_{t'}^B\|}. \quad (4)$$

If $\text{Sim}\left(s_t^A, s_{t'}^B\right) \geq \delta$, where δ is a predefined threshold, this pair of states is selected as a stitching point. Otherwise, new trajectories τ_A and τ_B are sampled, and the process repeats.

Once a valid stitching point is identified, two trajectories τ_A and τ_B are combined to construct a stitched trajectory τ_{stit}. We concatenate the segments of τ_B from time 0 through t' and the segments of τ_A from $t+1$ to its terminal point T:

$$\begin{aligned} \tau_{\text{stit}} = &\left\langle \left(s_0^B, a_0^B, r_0^B\right), \ldots, \left(s_{t'}^B, a_{t'}^B, r_{t'}^B\right) \right\rangle \\ &\oplus \left\langle \left(s_{t+1}^A, a_{t+1}^A, r_{t+1}^A\right), \ldots, \left(s_T^A, a_T^A, r_T^A\right) \right\rangle. \end{aligned} \quad (5)$$

The newly created stitched trajectories τ_{stit} are incorporated into the original offline dataset $\mathcal{D}$ to create an augmented dataset $\mathcal{D}_{\text{aug}}$. Finally, we then use $\mathcal{D}_{\text{aug}}$ to train the offline RL algorithms, as its increased diversity and broader coverage of treatment trajectories lead to improved performance.

As depicted in Figure 2(b), our *treatment stitching* process effectively combines beneficial segments from different trajectories. For instance, even though Trajectory B culminates in an adverse effect, its initial to intermediate transitions can still reflect meaningful clinical behaviors. By identifying a high-similarity state pair and merging the early segment of Trajectory B with the later segment of Trajectory A, we generate a new 'stitched trajectory' that inherits favorable characteristics from both. This enables the creation of clinically valid treatment trajectories that are not directly present in $\mathcal{D}$.

Despite its effectiveness, this direct stitching relies on the assumption that the offline dataset $\mathcal{D}$ is sufficiently large or homogeneous to contain enough similar intermediate states across trajectories for identifying valid stitching points. Unfortunately, in clinical practice, this assumption may not always hold. Clinical offline datasets can be extremely sparse or heterogeneous, making it challenging—or even impossible—to find valid stitching points due to the lack of similar intermediate states between trajectories. As a result, direct stitching alone may be inadequate in certain clinical settings.

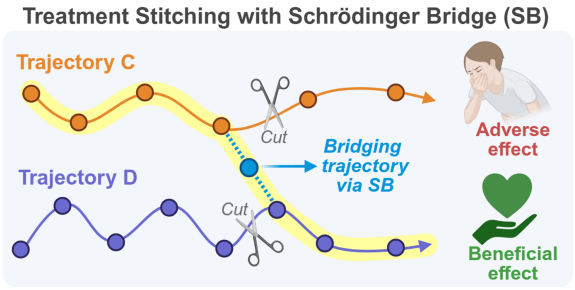

Figure 3: Overview of Schrödinger bridge for *TreatStitch*.

Schrödinger Bridge for Treatment Stitching

To tackle this challenge, we further introduce a novel mechanism that leverages the Schrödinger bridge method (Wang et al. 2021) to construct smooth and energy-efficient bridging trajectories that enable stitching even between dissimilar states. This enhancement broadens the applicability of our *TreatStitch* to sparse and heterogeneous clinical datasets.

Schrödinger Bridge. Originally, Erwin Schrödinger studied the problem of identifying the most probable and smooth stochastic trajectory that connects two given probability distributions (Schrödinger 1932). Nowadays, this is commonly referred to as the *Schrödinger bridge (SB) problem*, which is rooted in optimal transport (OT) theory (Léonard 2013).

Formally, given a starting distribution p_{start} and a target distribution p_{target}, and a reference stochastic process (e.g., Brownian motion), the goal is to find an optimal stochastic process $\mathbb{P}^*$ that transports samples from p_{start} to p_{target} over a finite time interval. This optimality is defined by minimizing the Kullback–Leibler (KL) divergence between the path measure of the target process $\mathbb{P}$ and the reference process $\mathbb{Q}$:

$$\mathbb{P}^* = \arg\min_{\mathbb{P}} \mathrm{KL}(\mathbb{P}\|\mathbb{Q}) \text{ s.t. } \mathbb{P}_0 = p_{\text{start}}, \mathbb{P}_T = p_{\text{target}}. \quad (6)$$

A key result is that the solution to this problem is characterized by the Schrödinger system (Caluya and Halder 2021), which consists of a pair of partial differential equations:

$$\begin{cases} \frac{\partial \Psi}{\partial t} = -\langle \nabla \Psi, f \rangle - \frac{1}{2} g^2 \Delta \Psi \\ \frac{\partial \hat{\Psi}}{\partial t} = -\nabla \cdot [\hat{\Psi} f] + \frac{1}{2} g^2 \Delta \hat{\Psi} \end{cases} \text{s.t.} \begin{cases} \Psi(\cdot, 0)\hat{\Psi}(\cdot, 0) = p_{\text{start}}, \\ \Psi(\cdot, T)\hat{\Psi}(\cdot, T) = p_{\text{target}}, \end{cases} \quad (7)$$

where Ψ and $\hat{\Psi}$ are the Schrödinger potentials, f denotes the drift vector field, and g is the scalar diffusion coefficient. A more detailed analysis of the SB problem is in Appendix E.

Problem Formulation. In terms of *TreatStitch*, we formulate the construction of bridging trajectories between dissimilar intermediate states as the SB problem. As shown in Figure 3, suppose we have two trajectories, Trajectory C (τ_C) and Trajectory D (τ_D), where no pair of intermediate states satisfies the similarity threshold δ. Our goal is to construct a bridging trajectory τ_{bridge} that connects a selected pair of intermediate states—$s_t^C \in \tau_C$ and $s_{t'}^D \in \tau_D$—identified as the most similar among all candidates despite still falling below the similarity threshold δ. This bridging trajectory τ_{bridge} serves to facilitate the stitching process between τ_C and τ_D.

Bridging State Generation. We first generate bridging states using the SB method. Specifically, we construct a neural network G_ϕ that learns the OT-based stochastic trajectory connecting two given probability distributions by solving the Schrödinger system as stated in Equation 7. Given the start state s_t^C and the target state $s_{t'}^D$, the goal of G_ϕ is to generate a sequence $s_{\text{bridge}} = \{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_K\}$ that ensures a smooth transition from s_t^C to $s_{t'}^D$. Moreover, the number of bridging states K is minimized to mitigate the risk of error accumulation resulting from extensive synthetic data generation. Due to space constraints, the detailed training objective and the step-by-step procedure for G_ϕ are provided in Appendix F.

Bridging Trajectory Completion. Once the sequence of s_{bridge} is obtained, we complete the bridging trajectory τ_{bridge} by inferring the corresponding actions and rewards for each transition. To this end, we construct two neural networks: an inverse dynamics model $I_\psi(a_t|s_t, s_{t+1})$ that predicts the action given a pair of successive states, and a reward prediction model $R_\rho(r_t|s_t, a_t)$ that estimates the reward based on a given state-action pair (Qing et al. 2025). These models are trained jointly using a supervised learning objective over trajectories sampled from $\mathcal{D}$. The combined loss function is:

$$\mathcal{L}(\psi, \rho) = \mathbb{E}_{\tau \sim \mathcal{D}}\Big[\|I_\psi(s_t, s_{t+1}) - a_t\|^2 + \|R_\rho(s_t, a_t) - r_t\|^2\Big]. \quad (8)$$

After training, we apply I_ψ and R_ρ to each adjacent pair of bridge states $(\tilde{s}_i, \tilde{s}_{i+1})$ to generate the corresponding action $\tilde{a}_i$ and reward $\tilde{r}_i$, thereby forming the bridging trajectory:

$$\tau_{\text{bridge}} = \langle (\tilde{s}_1, \tilde{a}_1, \tilde{r}_1), (\tilde{s}_2, \tilde{a}_2, \tilde{r}_2), \ldots, (\tilde{s}_K, \tilde{a}_K, \tilde{r}_K) \rangle. \quad (9)$$

Stitched Trajectory via Schrödinger Bridge Finally, we can construct an additional stitched trajectory via SB method $\tau_{\text{stit}}^{\text{SB}}$ by concatenating the initial segment of τ_C up to state s_t^C, the bridging trajectory τ_{bridge}, and the later segment of τ_D:

$$\begin{aligned} \tau_{\text{stit}}^{\text{SB}} = &\langle (s_0^C, a_0^C, r_0^C), \ldots, (s_t^C, a_t^C, r_t^C) \rangle \oplus \tau_{\text{bridge}} \\ &\oplus \langle (s_{t'}^D, a_{t'}^D, r_{t'}^D), \ldots, (s_T^D, a_T^D, r_T^D) \rangle \end{aligned} \quad (10)$$

Theoretical Justification for Treatment Stitching

We present the theoretical justification for the advantages of *TreatStitch*, showing that it preserves clinical validity by mitigating distribution shifts and avoiding OOD transitions.

Theorem 1. *Let $\mathcal{F} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ be the transition function defining the environment dynamics, and fix a norm $\|\cdot\|$ on $\mathcal{S}$. Suppose $\mathcal{F}$ is L-Lipschitz continuous in the state coordinate:*

$$\|\mathcal{F}(s, a) - \mathcal{F}(s', a)\| \le L\|s - s'\| \quad \forall\, s, s' \in \mathcal{S}, a \in \mathcal{A}. \quad (11)$$

Given the offline dataset $\mathcal{D} = \{\tau_i\}_{i=1}^N$, we construct the stitched trajectory $\tau_{stit} = \tau_B[0 : t'] \cup \tau_A[t + 1 : T]$, where the stitching point is chosen to satisfy a similarity condition, allowing a maximum difference of $1 - \delta$. Then, for every transition $(\bar{s}, a, \bar{s}')$ in τ_{stit}, there exists at least one transition (s, a, s') in $\mathcal{D}$ such that:

$$\|\bar{s} - s\| \le \sqrt{2(1 - \delta)} \ \wedge\ \|\bar{s}' - s'\| \le L\sqrt{2(1 - \delta)}. \quad (12)$$

Consequently, τ_{stit} remains within an $\mathcal{O}\big(L\sqrt{2(1 - \delta)}\big)$-neighborhood of transitions already in $\mathcal{D}$. In other words, our treatment stitching process stays within a small neighborhood of the original data support, reducing OOD shifts.

Method	Env1	Env2	Env3	Env4	Env5	Env6	Env7	Env8	Mean
CQL (Backbone)	59.65	57.03	47.25	53.82	54.35	54.68	55.59	52.55	54.37
+ GAN	32.99	50.20	38.72	29.20	16.82	2.99	29.48	20.09	27.56
+ DDPM	33.68	26.74	45.40	49.19	38.05	42.44	25.87	46.88	38.53
+ SynthER	33.09	25.90	40.74	31.29	44.62	53.25	28.84	49.83	38.45
+ GTA	61.75	56.55	43.26	53.17	47.60	56.96	57.55	47.82	53.08
+ ATraDiff	61.66	54.38	49.31	52.63	50.47	58.10	52.76	48.34	53.46
+ RTDiff	63.94	55.43	47.81	56.89	51.86	56.78	58.03	49.52	55.03
+ TreatStitch	65.15	**62.05**	50.57	58.22	**57.88**	56.98	59.09	**53.51**	57.93
+ TreatStitch w/ SB	**66.76**	60.32	**51.76**	**58.88**	55.83	**58.33**	**60.54**	52.03	**58.06**

Table 1: Performance comparison of each method using the CQL backbone on EpiCare under the **full data** setting.

Method	Env1	Env2	Env3	Env4	Env5	Env6	Env7	Env8	Mean
CQL (Backbone)	20.14	18.48	12.99	16.18	9.03	20.26	10.63	18.60	15.79
+ GAN	21.86	14.72	1.14	10.84	12.19	6.43	2.57	6.72	9.56
+ DDPM	13.53	15.48	17.18	9.99	16.94	12.95	5.60	5.17	12.11
+ SynthER	11.13	23.43	13.16	6.14	12.67	16.53	4.36	14.19	12.70
+ GTA	19.93	20.83	18.27	24.21	12.23	20.72	8.52	18.57	17.91
+ ATraDiff	13.29	21.53	14.24	24.49	15.53	25.10	13.65	13.43	17.66
+ RTDiff	23.27	20.91	17.35	25.38	14.80	21.47	12.93	24.57	20.09
+ TreatStitch	42.22	43.21	30.29	33.63	34.78	34.68	36.03	30.19	35.63
+ TreatStitch w/ SB	**48.47**	**47.36**	**34.00**	**36.94**	**37.10**	**36.65**	**41.43**	**38.64**	**40.07**

Table 2: Performance comparison of each method using the CQL backbone on EpiCare under the **restricted data** setting

Experimental Results and Discussion

Datasets. To evaluate the performance of our *TreatStitch*, we utilized the EpiCare benchmark (Hargrave, Spaeth, and Grosenick 2024), which is one of the most recent and comprehensive benchmarks for medical treatment evaluation. In particular, we selected this benchmark for our main experiments due to its appeal as a generalized benchmark, which makes it well-suited for longitudinal clinical settings without being limited to a specific disease. The EpiCare benchmark consists of datasets containing $2^{17} = 131,072$ episodes for each of 8 distinct environment settings. As described in the benchmark, these datasets can be interpreted as being drawn from sequential treatment trials for 8 distinct diseases. In addition, we also evaluated our framework on the MIMIC-III database (Johnson et al. 2016), which contains real-world electronic health record datasets. Specifically, we followed the experimental setup of prior work (Luo et al. 2024) that focuses on sepsis treatment in the intensive care unit (ICU).

Setup. We closely followed the experimental protocol outlined in the EpiCare benchmark for our main experiments. Specifically, we used the average cumulative reward as the evaluation metric. For reward design, we assigned $+64$ for remission, -64 for adverse events, and a penalty in the range of $[-1, -4]$ to reflect treatment costs. We set the similarity threshold $\delta = 0.95$ to ensure clinically valid stitching while maintaining sufficient synthetic samples. For the MIMIC-III dataset, we also followed the experimental setup established by prior work (Luo et al. 2024). We evaluated performance using root mean squared error (MSE) for both IV fluid treatment (MSE_{iv}) and vasopressor treatment (MSE_{va}), along with weighted importance sampling (WIS) and doubly robust (DR). In terms of reward design, the sequential organ failure assessment (SOFA) score (Jones, Trzeciak, and Kline 2009) was included to measure the severity of organ dysfunction, and the national early warning score 2 (NEWS2) (Inada-Kim and Nsutebu 2018) was used to estimate the risk of mortality. Further details are provided in Appendix G.

Competing methods. In this study, we used CQL (Kumar et al. 2020) as the main backbone offline RL algorithm due to its superior performance compared to other methods (see Appendix B) and its widespread adoption as a backbone in various prior works (Kondrup et al. 2023; Cai et al. 2023). Since our *TreatStitch* is a data augmentation framework, we compared it against other data augmentation methods in RL domain. First, we considered established generative models, including generative adversarial network (GAN) (Goodfellow et al. 2014) and diffusion models such as DDPM (Ho, Jain, and Abbeel 2020). Specifically, we trained GAN and DDPM to generate synthetic treatment trajectories from scratch, which were then employed to augment the original dataset. Next, we incorporated SynthER (Lu et al. 2023), which leverages diffusion models to augment synthetic data and enhances training through synthetic experience replay. We also included GTA (Lee et al. 2024), which extends SynthER by augmenting synthetic trajectories to be both high-reward and dynamically consistent. Additionally, we evaluated ATraDiff (Yang and Wang 2024), which uses a coarse-to-fine strategy to efficiently generate synthetic trajectories. Finally, we included RTDiff (Yang and Wang 2025), which is a state-of-the-art method that synthesizes trajectories in the reverse direction for more effective data augmentation.

Full Data Results. As shown in Table 1, we evaluated the performance of each method under the full data setting using 131,072 episodes. Both GAN and DDPM significantly underperformed compared to the backbone, highlighting the risks of generating synthetic trajectories from scratch. This drop in performance is consistent with findings from a prior study (Shumailov et al. 2024) that describe model collapse, where generative models produce erroneous and low-quality synthetic data that leads to performance decline. Among the competing methods, RTDiff was the only method to show a performance improvement. However, this gain was limited, because RTDiff was designed for general RL tasks without considering clinical validity. In contrast, our *TreatStitch* demonstrated superior performance by leveraging the treatment stitching that preserves clinical validity. Rather than generating synthetic data from scratch, *TreatStitch* creates clinically valid new trajectories by intelligently combining segments from existing trajectories. The extended version, *TreatStitch w/ SB*, achieved slightly better performance than *TreatStitch*. However, this improvement was marginal, likely because the full data setting already provided abundant valid stitching points, thereby limiting additional gains from SB.

Method	SOFA (↓)				NEWS2 (↓)			
	MSE_{iv}	MSE_{va}	WIS	DR	MSE_{iv}	MSE_{va}	WIS	DR
CQL (Backbone)	611.30	0.39	13.27	-0.37	566.64	0.33	-4.70	-0.69
+ GTA	634.81	0.42	14.45	-0.39	574.20	0.37	-5.21	-0.73
+ ATraDiff	607.75	0.39	13.53	-0.37	552.32	0.32	-4.54	-0.66
+ RTDiff	601.77	0.32	12.92	-0.34	550.06	0.28	-4.13	-0.62
+ TreatStitch	589.59	**0.28**	11.10	-0.31	529.48	0.25	-3.60	**-0.56**
+ TreatStitch w/ SB	**578.90**	**0.28**	**10.86**	**-0.30**	**526.71**	**0.24**	**-3.33**	**-0.56**

Table 3: Performance comparison of each method using the CQL backbone on the real-world **MIMIC-III** dataset.

Restricted Data Results. Since certain clinical settings often suffer from limited data availability, we conducted additional experiments under the restricted data setting, using only $2^{10} = 1024$ episodes—128 times fewer than in the full data setting. This setting reflects scenarios where only limited patient treatment data is available, resulting in a sparse offline dataset. As shown in Table 2, both GAN and DDPM again demonstrated poor performance, confirming their fundamental limitations. Interestingly, other data augmentation methods—including GTA, ATraDiff, and RTDiff—showed modest performance improvements compared to the backbone. This contrasts with their performance in the full data setting and suggests that when data is extremely sparse, RL-specific augmentations methods can still offer some benefit by diversifying the training distribution. However, these improvements still remain limited, as these methods do not account for clinical validity in their design. In contrast, our *TreatStitch* achieved significantly better performance by preserving clinical validity. More importantly, *TreatStitch w/ SB* exhibited substantial additional gains under this restricted data setting. This is likely because, in the sparse offline dataset, the number of valid stitching points between trajectories is limited. Our SB method addresses this challenge by constructing bridging trajectories that enable stitching even between dissimilar intermediate states, thereby increasing the availability of viable stitching points and further enhancing model performance through additional data augmentation. Moreover, to assess the versatility and generalizability of *TreatStitch*, we integrated it with other offline RL models beyond CQL. The corresponding results are in Appendix C.

Real-World Data Results. We also evaluated our framework on the MIMIC-III dataset, which includes real-world clinical data on sepsis treatment in the ICU. We compared our *TreatStitch* against strong competing methods, including GTA, ATraDiff, and RTDiff. As demonstrated in Table 3, our *TreatStitch* framework consistently outperformed all competing methods, even when evaluated using clinically meaningful metrics such as SOFA and NEWS2, alongside standard off-policy evaluation metrics like WIS and DR. These results highlight the practical applicability and robustness of our framework when applied to real-world clinical data beyond benchmark settings. Additionally, *TreatStitch w/ SB* achieved a modest further performance gain, suggesting the potential benefits of SB method in real-world clinical data.

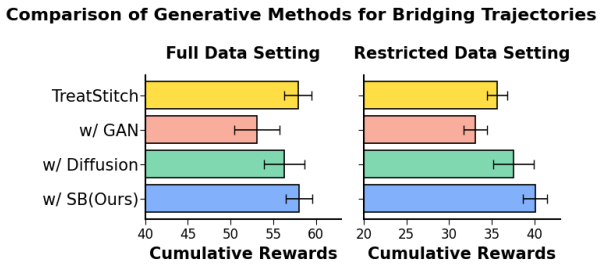

Figure 4: Comparison of various generative methods.

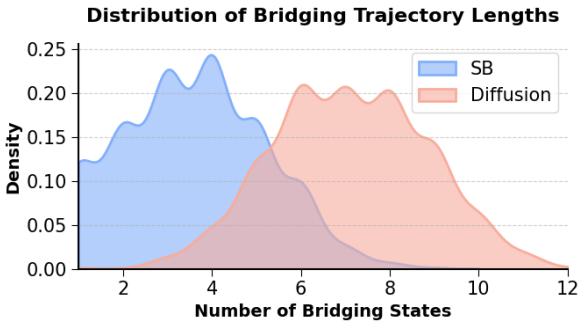

Figure 5: Distribution of bridging trajectory lengths.

Analysis of Bridging Trajectories. To analyze the efficacy of our SB method, we explored alternative generative methods for constructing bridging trajectories. Specifically, we developed two additional variants: *TreatStitch w/ GAN* and *TreatStitch w/ Diffusion*, the latter drawing inspiration from prior work (Li et al. 2024). As in Figure 4, the GAN method showed a significant performance drop in both full and restricted data settings, reaffirming the fundamental limitations of GAN in generating high-quality treatment data. While the diffusion method yielded comparable results, our SB method outperformed both methods across all settings. To provide quantitative evidence, we compared the distribution of bridging trajectory lengths generated by SB and diffusion methods. Specifically, we counted the number of generated bridging states for each method and visualized their probability distributions using kernel density estimation. As in Figure 5, our SB method displayed a pronounced peak at shorter lengths, indicating it generated shorter bridging trajectories than the diffusion method. This can be attributed to the fact that SB is based on OT theory, which aims to identify the smooth and energy-efficient trajectories between states.

Conclusion

In this work, we propose *TreatStitch*, a novel data augmentation framework designed to enhance offline RL for adaptive treatment strategies (ATS) in clinical settings. Unlike traditional methods that synthesize data from scratch, *TreatStitch* generates clinically valid synthetic trajectories by 'stitching' segments of existing treatment data. We empirically validate the efficacy of *TreatStitch* on multiple datasets, and theoretically demonstrate its ability to mitigate OOD transitions.

Acknowledgements

This work was supported by the Ministry of Science and ICT (MSIT), Korea, through the Institute of Information & Communications Technology Planning & Evaluation (IITP) [Artificial Intelligence Graduate School Program at Korea University, No. RS-2019-II190079]; the National Research Foundation of Korea (NRF) [No. RS-2023-00212498]; and the Korea Health Industry Development Institute (KHIDI) under the Federated Learning-based Drug Discovery Acceleration Project (K-MELLODDY) [No. RS-2025-16066488] and the Frailty Zero Project [No. RS-2025-25455839].

Contribution Statement: Dong-Hee Shin is the first author, and Tae-Eui Kam is the corresponding author. Deok-Joong Lee contributed to the discussion, and Young-Han Son assisted with proofreading. All authors have read and approved the final version of the manuscript.

References

Agarwal, R.; Schuurmans, D.; and Norouzi, M. 2020. An optimistic perspective on offline reinforcement learning. In *International conference on machine learning*. PMLR.

An, G.; Moon, S.; Kim, J.-H.; and Song, H. O. 2021. Uncertainty-based offline reinforcement learning with diversified q-ensemble. *Advances in neural information processing systems*, 34: 7436–7447.

Bauer, A.; Trapp, S.; Stenger, M.; Leppich, R.; Kounev, S.; Leznik, M.; Chard, K.; and Foster, I. 2024. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*.

Cai, X.; Chen, J.; Zhu, Y.; Wang, B.; and Yao, Y. 2023. Towards real-world applications of personalized anesthesia using policy Constraint Q Learning for Propofol Infusion Control. *IEEE Journal of Biomedical and Health Informatics*.

Caluya, K. F.; and Halder, A. 2021. Wasserstein proximal algorithms for the Schrödinger bridge problem: Density control with nonlinear drift. *IEEE Transactions on Automatic Control*, 67(3): 1163–1178.

Chen, T.; Liu, G.-H.; and Theodorou, E. A. 2021. Likelihood training of Schrödinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*.

De Bortoli, V.; Thornton, J.; Heng, J.; and Doucet, A. 2021. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34: 17695–17709.

Eghbali, N.; Alhanai, T.; and Ghassemi, M. M. 2025. Distribution-Free Uncertainty Quantification in Mechanical Ventilation Treatment: A Conformal Deep Q-Learning Framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27960–27968.

Fatemi, M.; Wu, M.; Petch, J.; Nelson, W.; Connolly, S. J.; Benz, A.; Carnicelli, A.; and Ghassemi, M. 2022. Semi-markov offline reinforcement learning for healthcare. In *Conference on Health, Inference, and Learning*, 119–137. PMLR.

Fujimoto, S.; and Gu, S. S. 2021. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34: 20132–20145.

Giordano, C.; Brennan, M.; Mohamed, B.; Rashidi, P.; Modave, F.; and Tighe, P. 2021. Accessing artificial intelligence for clinical decision-making. *Frontiers in digital health*, 3.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hargrave, M.; Spaeth, A.; and Grosenick, L. 2024. EpiCare: A Reinforcement Learning Benchmark for Dynamic Treatment Regimes. In *Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Inada-Kim, M.; and Nsutebu, E. 2018. NEWS 2: an opportunity to standardise the management of deterioration and sepsis. *Bmj*, 360.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Jones, A. E.; Trzeciak, S.; and Kline, J. A. 2009. The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Critical care medicine*, 37(5): 1649–1654.

Kondrup, F.; Jiralerspong, T.; Lau, E.; de Lara, N.; Shkrob, J.; Tran, M. D.; Precup, D.; and Basu, S. 2023. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15696–15702.

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

Kumar, A.; Zhou, A.; Tucker, G.; and Levine, S. 2020. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33.

Lee, J.; Yun, S.; Yun, T.; and Park, J. 2024. GTA: Generative Trajectory Augmentation with Guidance for Offline Reinforcement Learning. *Advances in Neural Information Processing Systems*.

Léonard, C. 2013. A survey of the Schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*.

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

Li, G.; Shan, Y.; Zhu, Z.; Long, T.; and Zhang, W. 2024. DiffStitch: Boosting Offline Reinforcement Learning with Diffusion-based Trajectory Stitching. In *International Conference on Machine Learning*, 28597–28609. PMLR.

Liu, Y.; Logan, B.; Liu, N.; Xu, Z.; Tang, J.; and Wang, Y. 2017. Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *2017 IEEE international conference on healthcare informatics (ICHI)*. IEEE.

Lu, C.; Ball, P.; Teh, Y. W.; and Parker-Holder, J. 2023. Synthetic experience replay. *Advances in Neural Information Processing Systems*, 36: 46323–46344.

Luo, Z.; Pan, Y.; Watkinson, P.; and Zhu, T. 2024. Position: reinforcement learning in dynamic treatment regimes needs critical reexamination. In *International conference on machine learning*. PMLR.

Mao, X. 2015. The truncated Euler–Maruyama method for stochastic differential equations. *Journal of Computational and Applied Mathematics*, 290: 370–384.

Murphy, S. A. 2005. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10): 1455–1481.

Nair, A.; Gupta, A.; Dalal, M.; and Levine, S. 2020. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qing, Y.; Chen, S.; Chi, Y.; Liu, S.; Lin, S.; and Zou, C. 2025. BiTrajDiff: Bidirectional Trajectory Generation with Diffusion Models for Offline Reinforcement Learning. *arXiv preprint arXiv:2506.05762*.

Schaul, T.; Quan, J.; Antonoglou, I.; and Silver, D. 2015. Prioritized experience replay. In *International Conference on Learning Representations*.

Schrödinger, E. 1932. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, 269–310.

Shaheen, M. Y. 2021. Applications of Artificial Intelligence (AI) in healthcare: A review. *ScienceOpen Preprints*.

Shi, Y.; De Bortoli, V.; Campbell, A.; and Doucet, A. 2023. Diffusion Schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36: 62183–62223.

Shin, D.-H.; Ko, D.-H.; Han, J.-W.; and Kam, T.-E. 2022. Evolutionary reinforcement learning for automated hyperparameter optimization in EEG classification. In *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*, 1–5. IEEE.

Shin, D.-H.; Lee, D.-J.; Han, J.-W.; Son, Y.-H.; and Kam, T.-E. 2025a. Population-based evolutionary search for joint hyperparameter and architecture optimization in brain-computer interface. *Expert Systems with Applications*, 264: 125832.

Shin, D.-H.; Son, Y.-H.; and Kam, T.-E. 2025. Sharpness-aware minimization with physics-informed regularizations for predicting semiconductor material properties in molecular dynamics. *Chemometrics and Intelligent Laboratory Systems*, 105511.

Shin, D.-H.; Son, Y.-H.; Kim, J.-M.; Ahn, H.-J.; Seo, J.-H.; Ji, C.-H.; Han, J.-W.; Lee, B.-J.; Won, D.-O.; and Kam, T.-E. 2024a. MARS: Multiagent reinforcement learning for spatial–spectral and temporal feature selection in EEG-based BCI. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Shin, D.-H.; Son, Y.-H.; Lee, D.-J.; Han, J.-W.; and Kam, T.-E. 2024b. Dynamic many-objective molecular optimization: Unfolding complexity with objective decomposition and progressive optimization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 6026–6034.

Shin, D.-H.; Son, Y.-H.; Lee, H. J.; Lee, D.-J.; and Kam, T.-E. 2025b. Offline Model-based Optimization for Real-World Molecular Discovery. In *International Conference on Machine Learning*. PMLR.

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1): 1–48.

Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; and Gal, Y. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Theodoropoulos, P.; Komianos, N.; Pacelli, V.; Liu, G.-H.; and Theodorou, E. A. 2025. Feedback Schrödinger bridge matching. In *International Conference on Learning Representations*.

Torabi, F.; Warnell, G.; and Stone, P. 2018. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*.

Wang, G.; Jiao, Y.; Xu, Q.; Wang, Y.; and Yang, C. 2021. Deep generative learning via Schrödinger bridge. In *International conference on machine learning*. PMLR.

Wang, L.; Yu, W.; He, X.; Cheng, W.; Ren, M. R.; Wang, W.; Zong, B.; Chen, H.; and Zha, H. 2020. Adversarial cooperative imitation learning for dynamic treatment regimes. In *Proceedings of The Web Conference 2020*, 1785–1795.

Wang, L.; Zhang, W.; He, X.; and Zha, H. 2018. 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*, 2447–2456.

Yang, Q.; and Wang, Y. 2025. Rtdiff: Reverse trajectory synthesis via diffusion for offline reinforcement learning. In *International Conference on Learning Representations*.

Yang, Q.; and Wang, Y.-X. 2024. ATraDiff: Accelerating Online Reinforcement Learning with Imaginary Trajectories. In *International Conference on Machine Learning*, 56485–56500. PMLR.

Yu, C.; Liu, J.; Nemati, S.; and Yin, G. 2021. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1): 1–36.

Zhang, J.; and Bareinboim, E. 2019. Near-optimal reinforcement learning in dynamic treatment regimes. *Advances in Neural Information Processing Systems*, 32.

Zhang, X. N.; Pu, Y.; Kawamura, Y.; Loza, A.; Bengio, Y.; Shung, D.; and Tong, A. 2024. Trajectory flow matching with applications to clinical time series modelling. *Advances in Neural Information Processing Systems*, 37.