# Off-Policy Selection for Initiating Human-Centric Experimental Design

**Ge Gao**[*§] **Xi Yang**[†] **Qitong Gao**[‡] **Song Ju**[§] **Miroslav Pajic**[‡] **Min Chi**[§]

## Abstract

In human-centric tasks such as healthcare and education, the *heterogeneity* among patients and students necessitates personalized treatments and instructional interventions. While reinforcement learning (RL) has been utilized in those tasks, off-policy selection (OPS) is pivotal to close the loop by offline evaluating and selecting policies without online interactions, yet current OPS methods often overlook the heterogeneity among participants. Our work is centered on resolving a *pivotal challenge* in human-centric systems (HCSs): ***how to select a policy to deploy when a new participant joining the cohort, without having access to any prior offline data collected over the participant?*** We introduce First-Glance Off-Policy Selection (FPS), a novel approach that systematically addresses participant heterogeneity through sub-group segmentation and tailored OPS criteria to each sub-group. By grouping individuals with similar traits, FPS facilitates personalized policy selection aligned with unique characteristics of each participant or group of participants. FPS is evaluated via two important but challenging applications, intelligent tutoring systems and a healthcare application for sepsis treatment and intervention. FPS presents significant advancement in enhancing learning outcomes of students and in-hospital care outcomes.

## 1 Introduction

Human-centric systems (HCSs), *e.g.*, used in healthcare facilities [53, 45, 16] and intelligent education (IE) [5, 26, 68], have widely employed reinforcement learning (RL) to enhance user experience by improving outcomes of disease treatment, knowledge gaining, etc. Specifically, RL has been used in healthcare to automate treatment procedures [53], or in IE that can induce policies automatically adapting difficulties of course materials and helping students to setup and refine study plans to improve learning outcomes [34, 84]. Though various existing offline RL methods can be adopted [19, 29, 4] for policy optimization, validation of policies' performance is often conducted by online testing [61, 74, 69, 10]. Given the long testing horizon (*e.g.*, several years, or semesters, in healthcare, and IE, respectively) and the high cost of recruiting participants, online testing is considered exceedingly time- and resource-consuming, and sometimes could even be hindered by protocols overseeing human involved experiments, *e.g.*, performance and safety justifications need to be provided before new medical device controllers can be tested on patients [51].

Recently, off-policy evaluation (OPE) methods have been proposed to tackle such challenges by estimating the performance of target (evaluation) RL policies with offline data, which only requires the trajectories collected over behavioral polices given *a priori*; similarly, off-policy selection (OPS) targets to determine the most promising policies, out of the ones trained with different algorithms or hyper-parameter sets, that can be used for online deployment [3, 9, 46, 76, 81]. However, most existing OPS and OPE methods are designed in the context of homogenic agents, such as in robotics

[*]Stanford University. The work was done at North Carolina State University. Contact: `gegao@stanford.edu, mchi@ncsu.edu`.

[†]IBM Research

[‡]Duke University

[§]North Carolina State University

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

or games, where characteristics of the agents can be captured by their specifications, which are in general assumed fully known (*e.g.*, degree of freedom, angular constraint of each joint).

**The pivotal challenge in OPE/OPS for HCSs.** In contrast, in HCSs, the participants can have highly diverse backgrounds, where each person may be associated with unique underlying characteristics that are not straightforward to be captured individually; due to the partial observability of participants' mind states and the limited size of the cohort that can be recruited for experiments with HCSs. For example, patients participated in healthcare research studies could have different health/disease records, while the students using an intelligent tutoring system in IE may have different mindsets toward studying the course. As a result, the optimal criteria for selecting the policy to be deployed to each participant can vary, and, more importantly, it would be *intractable for existing OPS/OPE frameworks to determine what the policy selection criteria would be for a new participant who just joined the cohort.* Consequently, there lacks a framework that can resolve the *pivotal challenge* in facilitating real-world HCSs – *how to select a policy to deploy when a new participant joining the cohort, without having access to any prior offline data collected over the participant?*

In this work, we introduce **F**irst-glance off-**P**olicy **S**election (**FPS**), to address the problem of determining the OPS criteria needed for each new participant joining the cohort (*i.e.*, at $t = 0$ only, or without using information obtained from $t >= 1$ onwards), assuming that we have access to offline trajectories for a small batch of participants *a priori*, *i.e.*, the offline data. Specifically, it first partitions the participants from the offline dataset into sub-groups, clustering together the ones pertaining to similar behaviors. Then, an unbiased value function estimator, with bounded variance, is developed to determine the policy selection criteria for each sub-group. At last, when new participants join, they will be recommended with policies selected according to the sub-groups they fall within. Note that FPS is distinguished from typical off-policy selection (OPS) setup in the sense that, the major goal of prior OPS approaches is to select the best policy over the entire population, while FPS aims to decide the best policy for each student who arrives to the HCS on-the-fly, leveraging the information observed at the initial step ($t = 0$) only.

The key contributions of this work are summarized as follows: (*i*) We introduce the FPS framework which is critical for closing the gap between OPS and applications pertaining to HCSs, *i.e.*, selecting the policy that would maximize the gain of the new participants at the point of joining a cohort. To the best of our knowledge, this is the first framework that considers the new participant arrival's problem in the context of OPS in HCSs. (*ii*) We conduct extensive experiments to evaluate FPS in a ***real-world IE system***, with 1,288 students participating over 5 years. Results have shown that, with the help of FPS, it improved the learning outcomes by $208\%$ compared to policy selection criteria hand-crafted by instructors. Moreover, it leads to $136\%$ increased outcome compared to policies selected by existing OPS methods. (*iii*) FPS is also evaluated against ***an important healthcare application***, *i.e.*, septic shock treatment [45, 48, 53], where it can accurately identifying the best treatment policies to be deployed to incoming patients, and outperforms existing OPS methods.

## 2 First-Glance Off-Policy Selection (FPS)

In this section, we introduce the FPS method, which determines the policy to be deployed to new participants that join an existing cohort, conditioned only on their initial states. Specifically, the participants pertaining to the offline dataset are partitioned into sub-groups according to their past behavior. Then, a variational auto-encoding (VAE) model is used to generate synthetic trajectories for each sub-group, augmenting the dataset and improving the state-action coverage. Moreover, an unbiased value function estimator, with bounded variance, is developed to determine the policy selection criteria for each sub-group. At last, when new participants join, they will be recommended with the policies conditioned on the sub-groups they fall within respectively. We start with a sub-section that introduces the problem formulation formally.

### 2.1 Problem Formulation

The HCS environment is formulated as a human-centric Markov decision process (HC-MDP), which is a 7-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{S}_0, R, \mathcal{I}, \gamma)$. Specifically, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ defines transition dynamics from the current state and action to the next state, $\mathcal{S}_0$ defines the initial state distribution, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\mathcal{I}$ is the set of participants involved in the HCS, $\gamma \in (0, 1]$ is discount factor. Episodes are of finite horizon $T$. At each time-step $t$ in *online* policy deployment, the agent observes the state $s_t \in \mathcal{S}$ of the environment, then chooses an action $a_t \in \mathcal{A}$ following the *target (evaluation)* policy $\pi$. The environment accordingly provides a reward $r_t = R(s_t, a_t)$, and the agent observes the next state $s_{t+1}$ determined by $\mathcal{P}$. A

*trajectory* is denoted as $\tau_\pi^{(i)} = [\dots, (s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)}), \dots]_{t=1}^T$. Moreover, we consider having access to a historical trajectory set (*i.e.*, offline dataset) collected under a *behavioral* policy $\beta \neq \pi$, $\mathcal{D}_\beta = \{..., \tau_\beta^{(i)}, ...\}_{i=1}^N$, which consist of $N$ trajectories. We first make two assumptions in regards to the correspondence between trajectories and participants, and the initial state distribution for each participant, respectively.

**Assumption 2.1** (Trajectory-Participant Correspondence)**.** As a participant in human-centric experiments is in general unlikely to undergo exactly the same procedure more than once under the topic being studied, we assume that there exist a unique correspondence between each trajectory $\tau^{(i)}$ and the participant ($i \in \mathcal{I}$) from which the trajectory is logged.

We henceforth can use $i$ to refer to index either a trajectory from the offline dataset, or the corresponding participant, depending on the context.

**Assumption 2.2** (Independent Initial State Distributions)**.** The initial state of each trajectory $s_0^{(i)} \in \tau^{(i)}$, corresponding to a unique (the i-th) participant following from the assumption above, is sampled from an initial state distribution $\mathcal{S}_0(i)$ conditioned on i-th participant's characteristics and past records (*i.e.*, specific to the i-th trajectory), and is independent from all other $\mathcal{S}_0(j)$'s where $j \in [1, N] \backslash i$.[5]

The assumptions above reflect the scenarios that are specific to HCS – for example, a patient is unlikely to be prescribed the same surgery twice. Even if the patient has to undergo a follow-up surgery that is of the similar type (*e.g.*, mostly seen in trauma or orthopedics departments), the second time when the patient comes in he/she will start with a rather different initial state, since the pathology may have already been intervened as a result of the last visit. Consequently, one can treat such a visit as a new (synthetic) participant who just join and has the health record same as the one updated after the last visit. In other words, a *participant* being considered in this paper can be generalized, *e.g.*, to a *hospital visit*, or a *student* participating in a *specific course* supported by intelligent education (IE) systems, depending on the context. Moreover, assumption 2.2 directly follows from the philosophy illustrated in assumption 2.1 – the initial state of each trajectory depend on the corresponding participant's unique characteristics and historical records before joining the experiment/cohort, and can be considered mutually independent across all participants. Now we define the goal for FPS.

**Problem 2.3.** *The goal of FPS is to select the best policy $\pi$ from a set of* ***pre-trained*** *(candidate) policies $\mathbf{\Pi}$, $\pi \in \mathbf{\Pi}$, for each of the new participants $i' \in \{N+1, N+2, \dots\}$ joining (i.e., arriving at) the HCS with an observable initial state $s_0 \sim \mathcal{S}_0(i')$ (but the rest of the trajectory remain unobservable), that maximizes the expected accumulated return $V^\pi$, $\max_{\pi \in \mathbf{\Pi}} V^\pi$, over the full horizon $T$; here $V^\pi = \mathbb{E}_{s_0 \sim \mathcal{S}_0(i'), (s_{t>0}, a_{t>0}) \sim \rho^\pi, r \sim R}[\sum_{t=1}^T \gamma^{t-1} r_t | \pi]$, and $\rho^\pi$ is the state-action visitation distribution under $\pi$ from step $t = 1$ onwards.*

Note that the problem formulation here is different than the typical OPS/OPE setup used in existing works [23, 66, 7, 77, 79, 14, 30], as only the initial state $s_0$ is available for policy selection. Such a formulation is aligned with use cases under HCSs, *e.g.*, treatment plan needs to be laid out soon after a new patient is admitted to the intensive care unit (ICU) in medical centers. However, most indirect OPS methods such as importance sampling (IS) [52, 7] and doubly robust (DR) [23, 66] require the entire trajectory to be observed, in order to estimate $V^\pi$. Though direct methods like fitted-Q evaluation (FQE) [30] could be used as a workaround, they do not take into account the unique characteristics for each participant that plays a crucial role in HCS applications; results in Section 3 show that they in general underperform in the real-world IE experiment. To address both challenges, we introduce the FPS approach, starting with the sub-group partitioning step introduced below.

### 2.2 Sub-Group Partitioning

In this sub-section, we introduce the sub-group partitioning step that partition the participants in the offline dataset into sub-groups. Furthermore, value functions over all candidate policies $\pi \in \mathbf{\Pi}$ are learned respectively for each sub-group, to be leveraged as the OPS criteria for each sub-group.

The partitioning is performed over the initial state of each trajectory in the offline dataset, $\tau_\beta \in \mathcal{D}$. Given assumptions 2.1 and 2.2, and the fact that $\mathcal{S}_0(i)$'s in general only share limited support across

[5] Without loss of generality, in the rest of the paper, we use $\mathcal{S}_0$ to represent the marginal distribution of the initial states over all participants, while $\mathcal{S}_0(i)$ represents the distribution specific to the $i$-th participant.

participants (*i.e.*, every human has unique characteristics and past experience), such partitioning is essentially performed at per-participant level. Specifically, we consider partitioning the participants into $M$ sub-groups. Then for all sub-groups, $K_m$'s, in the set of sub-groups, $\mathcal{K} = \{K_1, \dots, K_M\}$, we have $\bigcup_{m=1}^{M} K_m = \mathcal{S}_0$ and $K_m \cap K_n = \emptyset, \forall m \neq n$. The total number of groups $M$ needed can be determined using silhouette scores [21]. Denote the partition function $k(\cdot) : \mathcal{S}_0 \to \mathcal{K}$. We then define the value function specific to each sub-group.

**Definition 2.4** (Value Function per Sub-group)**.** The value function over policy $\pi$, $V^{\pi}_{K_m}$, specific to the sub-group $K_m$, is the expected accumulative return over the initial states that correspond to the set of participants $\mathcal{I}_m = \{i | k(s_0^{(i)}) = K_m, i \in \mathcal{I}\}$ residing in the same sub-group. $V^{\pi}_{K_m} = \mathbb{E}_{s_0 \sim Unif(\{\mathcal{S}_0(i) | i \in \mathcal{I}_m)\},\, (s_{t>0}, a_{t>0}) \sim \rho^{\pi},\, r \sim R}[\sum_{t=1}^{T} \gamma^{t-1} r_t | \pi]$, with $s_0 \sim Unif(\{\mathcal{S}_0(i) | i \in \mathcal{I}_m\})$ representing that $s_0$ is sampled from a uniformly weighted mixture of distributions over $\{\mathcal{S}_0(i) | i \in \mathcal{I}_m\}$, pertaining to sub-group $K_m$.

The goal of sub-group partitioning is to learn the partition function $k(\cdot)$, such that the difference between the value of the best policy candidate, $\max_{\pi \in \mathbf{\Pi}} V^{\pi}_{K_m}$, and the value of the behavioral policy, $V^{\beta}_{K_m}$, is maximized for all participants $i \in \mathcal{I}$ and sub-groups $K_m \in \mathcal{K}$, *i.e.*,

$$\max_{k} \sum_{i \in \mathcal{I}} \left[ \left( \max_{\pi \in \mathbf{\Pi}} V^{\pi}_{K_m = k\left(s_0^{(i)}\right)} \right) - V^{\beta}_{K_m = k\left(s_0^{(i)}\right)} \right]. \tag{1}$$

The objective (1) is designed in the sense that participants may benefit more from the type of policies that fit better for their individual characteristics. For example, in IE, different candidate lecturing policies may be used toward prospective high- and low-performers respectively, as justified by the findings from our real-world IE experiment (centered around Figure 2 in Section 3.2). The value provided by different policies for a specific type of learners (*i.e.*, sub-group) could be different, measured by $V^{\pi}_{K_m} - V^{\beta}_{K_m}$ for all $\pi \in \mathbf{\Pi}$; here, $V^{\beta}_{K_m}$ captures the expected return from a instructor-designed, one-size-fit-all baseline (*i.e.*, behavioral) policy that is used to collect offline data [38, 68, 84, 11]. Then, it would be crucial to identify to which group each student belongs, as it can maximize the returns collected by each student throughout the horizon.

**Practical off-policy deployment to initiate human-centric experiments over the sub-group objective** (1)**.** Our focus is to select the policy that can possibly work the best for each incoming individual from a set of policy candidates that are pre-given, which is critical for RL policy deployment in real-world HCSs, considering online policy optimization can be high-stake. The overall practical off-policy deployment can be achieved using a two-step approach, *i.e.*, $(i)$ *pre-partitioning* with offline dataset, followed by $(ii)$ *deployment* upon observation of the initial states of arriving participants. Due to space limitation, the specific steps can be found in Appendix D.1.

**Proposition 2.5.** *Define the estimator $\hat{D}^{\pi,\beta}_{K_m}$ as, i.e.,*

$$\hat{D}^{\pi,\beta}_{K_m} = \frac{1}{|\mathcal{I}_m|} \sum_{i \in \mathcal{I}_m} \left( \omega_i \sum_{t=1}^{T} \gamma^{t-1} r_t^{(i)} - \sum_{t=1}^{T} \gamma^{t-1} r_t^{(i)} \right); \tag{2}$$

*here, $\mathcal{I}_m$ follows the definition above, which is the set of participants grouped in $K_m$; $\omega_i = \Pi_{t=1}^{T} \pi(a_t^{(i)} | s_t^{(i)}) / \beta(a_t^{(i)} | s_t^{(i)})$ is the IS weight for the i-th trajectory in the offline dataset; $s_t^{(i)}, a_t^{(i)}, r_t^{(i)}$ are the states, actions, rewards logged in the offline trajectory, respectively. Then, $\hat{D}^{\pi,\beta}_{K_m}$ is unbiased, with its variance bounded by, i.e.,*

$$Var(\hat{D}^{\beta,\pi}) \leq \left|\left| \sum_{t=1}^{T} \gamma^{t-1} r_t \right|\right|_{\infty}^{2} \left( \frac{1}{ESS} - \frac{1}{N} \right), \tag{3}$$

*with $ESS$ being the effective sample size [27].*

The proof of proposition 2.5 is derived from [25] and provided in Appendix D.2.

### 2.3 Trajectories Augmentation within Each Sub-Group

In HCSs, each sub-group may only contain a limited number of participants, due to the high cost of recruiting participants as well as time constraints in real-world experiments. For example, in

the IE experiment in Section 3, one sub-group only contains 45 students as a result from sub-group partitioning. Consequently, the overall offline trajectories within each group may cover limited visitations of the state and action spaces, and make the downstream policy selection task challenging [46]. Latent-model-based data augmentation has been commonly employed in previous offline RL [20, 31, 58, 14], to resolve similar issues. For this sub-section, we specifically consider the variational auto-encoder (VAE) architecture introduced in [14], as it is originally designed for offline setup as well. Now we briefly introduce the VAE setup, which can capture the underlying dynamics and generate synthetic offline trajectories to improve the state-action visitation coverage *within each subgroup*. Specifically, given the offline trajectories $\mathcal{T}_m$ specific to the subgroup $K_m$, the VAE consists of three major components, *i.e.*, $(i)$ the latent prior $p(z_0)$ that represents the distribution of the initial latent states over $\mathcal{T}_m$; $(ii)$ the encoder $q_\eta(z_t|s_{t-1}, a_{t-1}, s_t)$ that encodes the MDP transitions into the latent space; $(iii)$ the decoders $p_\xi(z_t|z_{t-1}, a_{t-1})$, $p_\xi(s_t|z_t)$, $p_\xi(r_{t-1}|z_t)$ that reconstructs new samples. The training objective is formulated as an evidence lower bound (ELBO) specifically derived for the architecture above. More details can be found in Appendix D.3. Consequently, for the trajectories in each subgroup, $\mathcal{T}_m$, the VAE can be trained to generate a set of synthetic samples, denoted as $\widehat{\mathcal{T}}_m$. In the Section 3.2, we further discuss and justify the need of trajectory augmentation through an real-world intelligent education (IE) experiment.

### 2.4 The FPS Algorithm

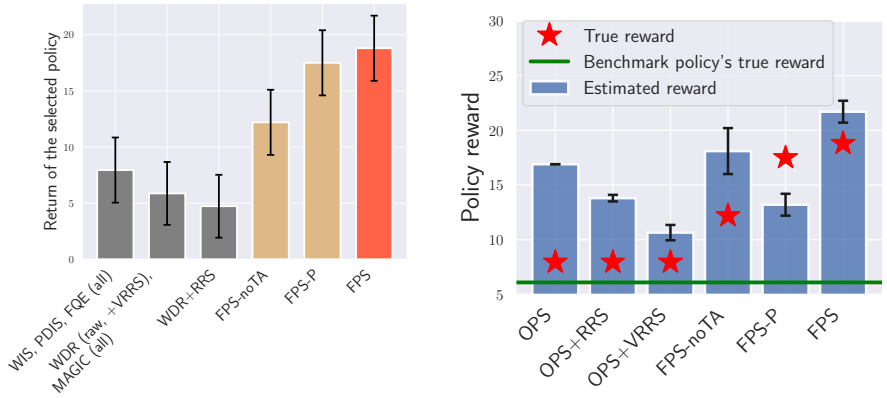
**Algorithm 1** FPS.

**Require:** A set of target policies $\mathbf{\Pi}$, offline dataset $\mathcal{D}$.
**Ensure:**
`// Training Phase.`
Calculate the number of subgroups $M$ needed for $\mathcal{D}$, using silhouette scores [21].
Obtain the sub-group partitioning $\mathbf{K} = \{K_1, \ldots, K_M\}$ following Section 2.2.
**for** each sub-group $K_m$ **do**
    Augment sub-group samples $\mathcal{T}_m$ with $\widehat{\mathcal{T}}_m$.
    Use the estimator in Proposition 2.5 to obtain $\hat{D}^{\pi,\beta}_{K_m}$ for all candidate target policies $\pi \in \mathbf{\Pi}$, over $\mathcal{T}_m \cup \widehat{\mathcal{T}_m}$.
    Select the best candidate target policy $\pi^*_m$ that maximizes $\hat{D}^{\pi,\beta}_{K_m}$ as the one to be deployed over $K_m$.
`// Deployment Phase.`
**while** the HCS receives the initial state $s_0$ from a new participant **do**
    Determine the sub-group $K_m$ for the new participant.
    Deploy to the participant the best candidate policy $\pi^*_m$ specific to sub-group $K_m$ .

The overall flow of the FPS framework is described in Algorithm 1. The training phase directly follow from the sub-sections above. Upon deployment, FPS can help HCSs monitor each arriving participant, determine the sub-group the participant falls within, and select the policy to be deployed according to the initial state. Such real-time adaptability is important for HCSs in practice, and is different from existing OPS works which in general assume either the full trajectories or population characteristics are known [25, 76, 82]. For example, in practical IE, students may start learning irregularly according to their own schedules, hence can create discrepancies in their start times. Such methods fall short in cases when selecting policies based on population or sub-group information in the upcoming semester – they requires the data from all arriving students are collected upfront, which would be unrealistic. Note that, to the best of our knowledge, we are the first work that formally consider the problem of sub-typing arriving participants, and FPS is the first approach that solves this practical problem by introducing a framework that can work with HCSs in the real-world.

## 3 Experiments

FPS is tested over two types of HCSs, *i.e.*, intelligent education (IE) and healthcare. Specifically, the ***real-world IE experiment*** involves 1,288 student participating in college entry-level probability course across 6 academic semesters. The goal is to use the data collected from the students of the first 5 semesters, to assign pre-trained RL lecturing policies to every student enrolled in the 6-th semester, in order to maximize their learning outcomes. The healthcare experiment targets for selecting pre-configured policies that can best treat patients with sepsis, over a simulated environment widely adopted in existing works [45, 46, 65, 36, 12].




(a) (b)

Figure 1: Analysis of main results from the real-world IE experiment. (a) Overall performance of the 6-th semester's student cohort. Methods that selected the same policy are merged in one bin, *i.e.*, `all` refers to all three variations (`raw, +RRS, +VRRS`) of the existing OPS baselines. (b) Estimated and true policy performance using each method. For *OPE*, *OPE+RRS*, *OPE+VRRS*, results with the least gap between estimated and true rewards among OPE methods (*i.e.*, WIS, FQE+RRS, and FQE+VRRS, respectively) are shown in the figure. True reward refers to the returns averaged over the cohort of the 6-th semester, obtained by deploying the policy selected for each student correspondingly.

### 3.1 Baselines

**Existing OPS/OPE.** The most straightforward approach to facilitate OPS in HCSs is to select policies via existing OPS/OPE methods, by choosing the candidate target policy $\pi \in \mathbf{\Pi}$ that achieves the maximum estimated return *over the entire offline dataset, i.e.*, indiscriminately across all potential sub-groups. Specifically, 6 commonly used OPE methods are considered, *i.e.*, Weighted IS (WIS) [52], Per-Decision IS (PDIS) [52], Fitted-Q Evaluation (FQE) [30], Weighted DR (WDR) [66], MAGIC [66], and Dual stationary DIstribution Correction Estimation (DualDICE) [44].

**Existing OPS/OPE with vanilla repeated random sampling (OPS+RRS).** We also compare FPS against a classic data augmentation method in order to evaluate the necessity of the VAE-based method introduced in Section 2.3 – *i.e.*, repeated random sampling (RRS) with replacement of the historical data to perform OPE. RSS has shown superior performance in some human-related tasks, such as disease treatment [46]. Specifically, all OPS/OPE methods considered above are applied to the RRS-augmented offline dataset, where the value of each candidate target policy is obtained by averaging over 20 sampling repetitions. However, note that RRS does not intrinsically consider the temporal relations among state-action transitions as captured by MDP.

**Existing OPS/OPE with VAE-based RRS (OPS+VRRS).** This baseline perform OPS with RRS on augmented samples resulted from the VAE introduced in Section 2.3, in order to allow RRS to consider MDP-typed transitions, hence improve state-action visitation coverage of the augmented dataset. This method can, to some extent, be interpreted as an ablation baseline of FPS, by removing the sub-group partitioning step (Section 2.2), and slightly tweaking the VAE-based offline dataset augmentation step (Section 2.3) such that it does not need any sub-group information. Specifically, we set the amount of augmented data identical to the amount of original historical data, *i.e.*, $|\widehat{\mathcal{T}}| = |\mathcal{T}| = N$, and RRS $N$ samples from both set $\widehat{\mathcal{T}} \cup \mathcal{T}$ to perform OPE. Final estimates are averaged results from 20 repeated sampling processes.

**FPS without trajectory augmentation (FPS-noTA).** This is the ablation baseline that completely removes from FPS the augmentation technique introduced in Section 2.3.

**FPS for the population (FPS-P).** We consider on additional ablation baseline that follows the same training steps as FPS (*i.e.*, steps 1-7 of Alg. 1), but rather select *a single policy* that is identified (by FPS) as the best for majority of the sub-groups, to be deployed to all participants. In other words, after training, FPS produces the mapping $h : \mathcal{K} \rightarrow \mathbf{\Pi}$, while FPS-P will always deploy to every arriving participant the policy that appears most frequently in the set $\{h(K_m) | K_m \in \mathcal{K}\}$.

### 3.2 The Real-World IE Experiment

The IE system has been integrated into a undergraduate-level introduction to probability and statistics course over 6 semesters, including a total of 1,288 student participants. This study has received approval from the Institutional Review Board (IRB) at the institution to ensure ethical compliance. Additionally, oversight is provided by a departmental committee, which is responsible for safeguarding the academic performance and privacy of the participants. In this educational context, each learning session revolves around a student's engagement with a set of 12 problems, with this period referred to as an "episode" (horizon $T = 12$). During each step, the IE system offers students three actions: independent work, utilizing hints, or directly receiving the complete solution (primarily for study purposes). The states space is constituted by by 140 features that have been meticulously extracted from the interaction logs by domain experts, which encompass various aspects of the students' activities, such as the time spent on each problem and the accuracy of their solutions. The learning outcome is issued as the environmental reward at the end of each episode (0 reward for all other steps), measured by the normalized learning gain (NLG) quantified using the scores received from two exams, *i.e.*, one taken before the student start using the system, and another after. Data collected from the first 5 semesters (over a lecturer-designed behavioral policy) are used to train FPS for selecting from a set of candidate policies to be deployed to each student in the cohort of the 6-th semester, including 3 pre-trained RL policies and 1 benchmark policy (whose performance benchmark the lower-bound of what could be tested with student participants). See Appendix A for the definition of NLG, details on pre-trained RL policies, and more.

**Main results.** Figure 1(a) presents students' performance under policies selected by different methods. Overall, FPS was the most effective policy selection leading to the greatest average student performance. The return difference between FPS and the two ablation, FPS-noTA and FPS-P, illustrate the importance of augmenting offline trajectories (as introduced in Section 2.3) and assign to arriving students policies that better fit the characteristics shared within their sub-groups, respectively. Moreover, most existing OPS/OPE methods tend to select sub-optimal policies that resulted in better learning gain than the benchmark policy. Note that we also observed that DualDICE could not distinguish the returns over all target policies; thus, it is unable to be used for policy selection in this experiment and we omit its results. It is also important to evaluate how accurate the value estimation $V^{\pi^*_\cdot}$ would be for the best candidate policy selected across all methods, over the arriving student cohort at the 6-th semester, as illustrated in Figure 1(b). FPS provided more accurate policy estimation by achieving the smallest error between true and estimated policy rewards. With VRRS, most OPS methods improved their policy estimation performance, which was benefited from the richer state-action visitation coverage provided by the synthetic samples generated by VRRS. However, even with such augmentations, existing OPS methods still chose sub-optimal policies, which justified the importance of considering participant-specific characteristics in HCSs, which is tackled by sub-group partitioning in FPS (Section 2.2).

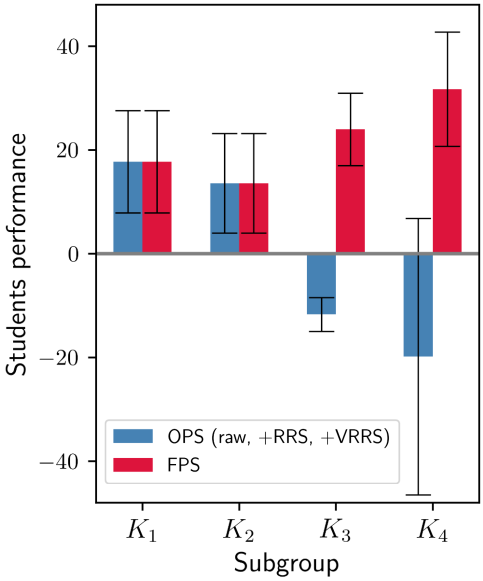



Figure 2: Performance of students (mean±se) over all four sub-groups under selected policies in the 6-th semester.

**More discussions.** For a more comprehensive understanding of student behaviors affected by the policy being deployed in IE, we further investigate how the sub-groups are partitioned and how the policies being assigned to each sub-group perform. Specifically, FPS identified four subgroups (*i.e.*, $K_1, K_2, K_3, K_4$) as a result of Section 2.2. Under the behavioral policy, the average NLG across all students is 0.9 with slight improvement after tutoring. Specifically, $K_1(N_{train} = 345, N_{test} = 30)$ and $K_2(N_{train} = 678, N_{test} = 92)$ achieved average NLG of 1.9 *[95% CI, 1.7, 2.1]*[6] and 0.7 *[95% CI, 0.6, 0.8]* under the behavioral policy, respectively. In the testing (6-th) semester, FPS

[6]CI stands for confidence interval.

Table 1: The absolute errors (AEs) and returns resulted from deploying to each patient the corresponding candidate policy selected by FPS against baselines, as well as the top-1 regret (regret@1) of the selected policy, averaged over 10 different simulation runs. Standard errors are rounded.

| | | FPS | WIS | PDIS | FQE | WDR | MAGIC |
|---|---|---|---|---|---|---|---|
| N=2,500 | AE | **0.026±0.00** | 0.054±0.00 | 0.109±0.00 | 0.070±0.01 | 8.281±0.00 | 5.681±0.00 |
| | Return | **0.132±0.02** | 0.121±0.01 | 0.121±0.01 | 0.129±0.00 | 0.121±0.01 | 0.129±0.00 |
| | Regret@1 | **0.042±0.01** | 0.066±0.00 | 0.066±0.00 | 0.106±0.09 | 0.066±0.00 | 0.106±0.09 |
| N=5,000 | AE | **0.006±0.00** | 0.046±0.00 | 0.082±0.00 | 0.073±0.01 | 3.443±0.01 | 3.238±0.01 |
| | Return | **0.149±0.01** | 0.123±0.01 | 0.123±0.01 | 0.121±0.01 | 0.123±0.01 | 0.123±0.01 |
| | Regret@1 | **0.020±0.00** | 0.050±0.00 | 0.050±0.00 | 0.208±0.13 | 0.050±0.00 | 0.050±0.00 |
| N=10,000 | AE | **0.022±0.00** | **0.022±0.00** | 0.097±0.00 | 0.105±0.00 | 0.995±0.01 | 1.210±0.01 |
| | Return | **0.130±0.00** | 0.129±0.00 | 0.121±0.00 | 0.121±0.00 | 0.121±0.00 | 0.121±0.00 |
| | Regret@1 | **0.016±0.00** | 0.019±0.00 | 0.029±0.01 | 0.029±0.01 | 0.029±0.01 | 0.029±0.01 |

constantly selected the best performing policy for students identified as sub-groups $K_1$ and $K_2$, with learning outcomes improvement quantified as 17.7 *[95% CI, -1.7, 37.1]* and 13.6 *[95% CI, -5.2, 32.4]* in terms of NLGs, respectively, as shown in Figure 2. This is on the same level achieved by the best possible baseline combinations worked for each student, regardless of the base OPS algorithm used, *i.e.*, the union of the best performance reported from the 18 baselines involving existing OPS methods (introduced in the first 3 paragraphs of Section 3.1) over each student. However, note that in reality one does not have access to such an oracle in terms of which 1 out of the 18 baseline methods would work well for each arriving student upfront (*i.e.*, at the beginning of the semester). In contrast, *FPS achieved the same level of performance without the need for an oracle*. Note that sub-groups $K_1$ and $K_2$, were less sensitive to target policies and achieved positive NLG in both training and testing semester. On the other hand, offline data over the behavioral policy showed that $K_3(N_{train} = 101, N_{test} = 12)$ and $K_4(N_{train} = 24, N_{test} = 6)$ are associated with negative average NLGs of -0.5 *[95% CI, 1.7, 2.1]* and -1.5 *[95% CI, -0.3, 1.2]*, respectively, which can be considered low performers. It is observed that students in sub-group $K_3$ performed kept moving rapidly among questions while working on the IE system, indicating that they were not seriously tackling any one of the questions; while participants in $K_4$ abused hints, but still made much more mistakes in the meantime. Figure 2 also presents the NLG of students from the low-performing subgroups $K_3$ and $K_4$ under policies selected by the best existing-OPS baselines (following the oracle as above) and FPS. Under FPS, both subgroups achieved significant improvement (average NLGs 24.0 *[95% CI, 10.2, 37.7]* and 31.7 *[95% CI, 10.1, 53.3]*, respectively) compared to students in historical semesters. However, the sub-optimal policy chosen by baselines had a negative effect on both sub-groups (average NLGs -11.7 *[95% CI, -18.1, -5.3]* and -19.9 *[95% CI, -72.1, 32.4]*, respectively); see Figure 1(a). Such an observation particularly justifies the need for personalizing the policies deployed to different type of participants (*i.e.*, students), especially for the sub-groups (*i.e.*, low-performers), since they can be more sensitive to policy selection. Based on the statistics reported above, FPS improved the NLG of students by 208% over the lecturer-designed behavioral policy, and by 136% over the union of the best performance achieved across existing-OPS-based baselines. Though the difference may not be large for potential high-performing sub-groups (e.g., $K_1$ & $K_2$), we observed that the baselines can even have a negative effect on some sub-groups (e.g.,$K_3$ & $K_4$), which is undesired in human-centric experiments. In empirical human-centric scenarios, such as education, the behavior policy is generally highly regularized by department committees strictly following guidance from human-centric experiments, such that the target policy would not be dramatically opposed to the behavior policy – suggesting the underlying assumption that the divergence between behavior and target policies could be intrinsically bounded. To this end, the FPS framework has the potential to facilitate fairness in RL-empowered HCSs in general – we have discussed this in details in Appendix A.5.

### 3.3 The Healthcare Experiment

In this experiment, we consider selecting the policy that can best treat sepsis for each patient in the ICU, leveraging the simulated environment introduced by [48], which has been widely adopted in existing works [22, 46, 65, 36, 12, 45]. Specifically, the state space is constituted by a binary indicator for diabetes, and four vital signs {heart rate, blood pressure, oxygen concentration, glucose level} that take values in a subset of {very high, high, normal, low, very low}; size of the state space is $|\mathcal{S}|$ = 1440. Actions are captured by combinations of three binary treatment options, {antibiotics, vasopressors, mechanical ventilation}, which lead to $|\mathcal{A}| = 2^3$. Three candidate target policies are considered and provided by [45], *i.e.*, $(i)$ without antibiotics (WOA) which does not administer antibiotics right

after the patient is admitted, $(ii)$ with antibiotics (WA) that always administer antibiotics once the patient is admitted, $(iii)$ an RL policy trained following policy iteration (PI). Note that as pointed by [45], the true returns of WA and PI are usually close, since antibiotics are in general helpful for treating sepsis, which is also observed in our experiment; see Table 1. Moreover, a simulated unrecorded comorbidities is applied to the cohort, capturing the uncertainties caused by patient's underlying diseases (or other characteristics), which could reduce the effects of the antibiotics being administered. See Appendix B for more details in regards to the environmental setup.

Given the simulated environment, we mainly consider using this experiment to evaluate the source of improvement brought in by the sub-group partitioning step (Section 2.2) in FPS. Specifically, multiple scaled offline datasets are generated, representing different degrees of the state-action visitation coverage – we vary the total number of trajectory $N$={2,500, 5,000, 10,000}, in lieu of performing trajectory augmentations for both FPS and existing OPS baselines. In other words, in this experiment, we consider the FPS without the VAE augmentation step introduced in Section 2.3, as well as the 6 original OPS baselines (without any RRS/VRRS) introduced in Section 3.1. We believe this setup would help isolate the source of improvements brought in by sub-group partitioning. The average absolute errors (AEs), in terms of OPE, and returns, in terms of OPS, resulted from deploying to each patient the corresponding candidate policy selected by FPS against baselines, are reported in Table 1. It can be observed that FPS achieved the lowest AE and highest return regardless of the size of the offline dataset. We additionally evaluate the top-1 regret (*i.e.*, regret@1) of the selected policy following FPS and baselines, which are also reported in Table 1. It can be observed that FPS achieved exceedingly low regrets compared to baselines. Both observations emphasize the effectiveness of the sub-group partitioning technique leveraged by FPS, as the environment does capture comorbidities as part of the participant characteristics. Moreover, the AEs and regrets of most methods decrease when the size of offline dataset increase, justifying that improved state-action visitation coverage provided by the offline trajectories is crucial for reducing estimation errors and improving policy selection outcomes (*i.e.*, the motivation of trajectory augmentation introduced in Section 2.3).

## 4 Related Works

**Off-policy selection (OPS).** OPS are typically approached via OPE in existing works, by estimating the expected return of target policies using historical data collected under a behavior policy. A variety of contemporary OPE methods has been proposed, which can be mainly divided into three categories [72]: (*i*) direct methods that directly estimate the value functions of the evaluation policy [44, 67, 79, 76], including but not limited to model-based estimators (MB) [17, 14, 15, 79], value-based estimators [30] such as Fitted Q Evaluation (FQE), and minimax estimators [35, 80, 70] such as DualDICE [77]; (*ii*) inverse propensity scoring, or indirect methods [52], such as Importance Sampling (IS) [7]; (*iii*) hybrid methods combine aspects of both inverse propensity scoring and direct methods [66], such as DR [23]. In practice, due to expensive online evaluations, researchers generally selected the policy with the highest estimated rewards via OPE. For example, Mandel et al. selected the policy with the maximum IS score to be deployed to an educational game [38]. Recently, some works focused on estimator selection or hyperparameter tuning in off-policy selection [46, 75, 63, 43, 28, 32, 65, 50]. However, retraining policies may not be feasible in HCSs as online data collection is time- and resource-consuming. More importantly, prior work generally selected policies without considering the characteristics of participants, while personalized policy is flavored towards the needs specific to HCSs.

**RL-empowered automation in HCSs.** In modern HCSs, RL has raised significant attention toward enhancing the experience of human participants. Previous studies have demonstrated that RL can induce IE policies [1, 38, 60, 73]. For example, Zhou et al. [84] applied hierarchical reinforcement learning (HRL) to improve students' normalized learning gain in a Discrete Mathematics course, and the HRL-induced policy was more effective than the Deep Q-Network induced policy. Similarly, in healthcare, RL has been used to synthesize policies that can adapt high-level treatment plans [53, 45, 36], or to control medical devices and surgical robotics from a more granular level [16, 37, 55]. Since online evaluation/testing is high-stake in practical HCSs, effective OPS methods are important in closing the loop, by significantly reducing the resources needed for online testing/deployment and preemptively justifying safety of the policies subject to be deployed.

## 5 Conclusion and Limitation

In this work, we introduced the FPS framework that facilitated policy selection in real-world HCSs; it tackled the off-policy deployment with new arrivals problem that is pivotal for RL policy deployment in HCSs. Unlike existing OPS methods, FPS customized the policy selection criteria for each sub-group respectively. FPS was tested in a real-world IE experiment and a simulated sepsis treatment environment, which significantly outperformed baselines. Though in the future it would be possible to extend FPS to a offline RL policy optimization framework, however, in this work we specifically focus on the OPS task in order to isolate the source of improvements brought in by sub-group partitioning and trajectory augmentation. Future avenues along the line of FPS also include deriving estimators (for Proposition 2.5) that allow bias-variance trade off, *e.g.*, by integrating WDR or MAGIC (to substitute the IS weights). Societal and broader impacts are discussed in Appendix C.

Compared to IE systems, HCSs in healthcare would be considered even more high-stakes, thus may further limit the options (i.e., policies) that are available to facilitate sub-grouping experiments, due to stricter clinical experimental guidelines. However, FPS has demonstrated its extraordinary capabilities over a real-world experiment that involved >1,200 participants with years of follow-ups, which showed its efficacy and scalability toward working with more challenging systems and larger cohorts as in healthcare, as the assumptions needed by FPS across these two systems would not change fundamentally. Moreover, potential underlying confounding may exist across the patient's initial states in healthcare, and it is also important to consider inputs from healthcare professionals during sub-grouping. As a result, one may further extend our framework toward such a direction, allowing it to function better in the healthcare domain.

**Acknowledgments**

This research was supported by the NSF Grants: Integrated Data-driven Technologies for Individualized Instruction in STEM Learning Environments (1726550), CAREER: Improving Adaptive Decision Making in Interactive Learning Environments (1651909), Generalizing Data-Driven Technologies to Improve Individualized STEM Instruction by Intelligent Tutors (2013502), NAIAD Award (2332744), and National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks, Grant CNS-2112562. This research was also sponsored in part by the AFOSR under award number FA9550-19-1-0169. We would also like to thank the anonymous reviewers for insightful comments that lead to improved paper presentations.

## References

[1] Mark Abdelshiheed, John Wesley Hostetter, Tiffany Barnes, and Min Chi. Leveraging deep reinforcement learning for metacognitive interventions across intelligent tutoring systems. In *International Conference on Artificial Intelligence in Education*, pages 291–303. Springer, 2023.

[2] Karo Castro-Wunsch, Alireza Ahadi, and Andrew Petersen. Evaluating neural networks as a method for identifying students in need of assistance. In *Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education*, pages 111–116, 2017.

[3] Yash Chandak, Shiv Shankar, Nathaniel Bastian, Bruno da Silva, Emma Brunskill, and Philip S Thomas. Off-policy evaluation for action-dependent non-stationary environments. *Advances in Neural Information Processing Systems*, 35:9217–9232, 2022.

[4] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[5] Min Chi, Kurt VanLehn, Diane Litman, and Pamela Jordan. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction*, 21(1):137–180, 2011.

[6] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

[7] Shayan Doroudi, Philip S Thomas, and Emma Brunskill. Importance sampling for fair policy selection. *Grantee Submission*, 2017.

[8] Salma Elmalaki. Fair-iot: Fairness-aware human-in-the-loop reinforcement learning for harnessing human variability in personalized iot. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, pages 119–132, 2021.

[9] Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, Cosmin Paduraru, et al. Benchmarks for deep off-policy evaluation. In *International Conference on Learning Representations*, 2021.

[10] Justin Fu, Mohammad Norouzi, Ofir Nachum, George Tucker, Ziyu Wang, Alexander Novikov, Mengjiao Yang, Michael R Zhang, Yutian Chen, Aviral Kumar, et al. Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*, 2021.

[11] Ge Gao, Qitong Gao, Xi Yang, Song Ju, Miroslav Pajic, and Min Chi. On trajectory augmentations for off-policy evaluation. In *The Twelfth International Conference on Learning Representations*, 2024.

[12] Ge Gao, Song Ju, Markel Sanz Ausin, and Min Chi. Hope: Human-centric off-policy evaluation for e-learning and healthcare. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 1504–1513, 2023.

[13] Ge Gao, Samiha Marwan, and Thomas W Price. Early performance prediction using interpretable patterns in programming process data. In *Proceedings of the 52nd ACM technical symposium on computer science education*, pages 342–348, 2021.

[14] Qitong Gao, Ge Gao, Min Chi, and Miroslav Pajic. Variational latent branching model for off-policy evaluation. In *The Eleventh International Conference on Learning Representations*, 2022.

[15] Qitong Gao, Ge Gao, Juncheng Dong, Vahid Tarokh, Min Chi, and Miroslav Pajic. Off-policy evaluation for human feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 9065–9091, 2023.

[16] Qitong Gao, Michael Naumann, Ilija Jovanov, Vuk Lesi, Karthik Kamaravelu, Warren M Grill, and Miroslav Pajic. Model-based design of closed loop deep brain stimulation controller using reinforcement learning. In *2020 ACM/IEEE 11th International Conference on Cyber-Physical Systems (ICCPS)*, pages 108–118. IEEE, 2020.

[17] Qitong Gao, Stephen L Schmidt, Karthik Kamaravelu, Dennis A Turner, Warren M Grill, and Miroslav Pajic. Offline policy evaluation for learning-based deep brain stimulation controllers. In *2022 ACM/IEEE 13th International Conference on Cyber-Physical Systems (ICCPS)*, pages 80–91. IEEE, 2022.

[18] Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021.

[19] 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.

[20] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

[21] David Hallac, Sagar Vare, Stephen Boyd, and Jure Leskovec. Toeplitz inverse covariance-based clustering of multivariate time series data. In *ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 215–223, 2017.

[22] Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR, 2021.

[23] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

[24] Song Ju. Identify critical pedagogical decisions through adversarial deep reinforcement learning. In *In: Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, 2019.

[25] Ramtin Keramati, Omer Gottesman, Leo Anthony Celi, Finale Doshi-Velez, and Emma Brunskill. Identification of subgroups with similar benefits in off-policy policy evaluation. In *Conference on Health, Inference, and Learning*, pages 397–410. PMLR, 2022.

[26] Kenneth R. Koedinger, John R. Anderson, William H. Hadley, and Mary A. Mark. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8(1):30–43, 1997.

[27] Augustine Kong. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348, 1992.

[28] Aviral Kumar, Anikait Singh, Stephen Tian, Chelsea Finn, and Sergey Levine. A workflow for offline model-free robotic reinforcement learning. In *Conference on Robot Learning*, pages 417–428. PMLR, 2022.

[29] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

[30] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

[31] Alex X Lee, Anusha Nagabandi, Pieter Abbeel, and Sergey Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *Advances in Neural Information Processing Systems*, 33:741–752, 2020.

[32] Jonathan Lee, George Tucker, Ofir Nachum, and Bo Dai. Model selection in batch policy optimization. In *International Conference on Machine Learning*, pages 12542–12569. PMLR, 2022.

[33] Bruno Lepri, Nuria Oliver, and Alex Pentland. Ethical machines: The human-centric use of artificial intelligence. *IScience*, 24(3), 2021.

[34] Evan Liu, Moritz Stephan, Allen Nie, Chris Piech, Emma Brunskill, and Chelsea Finn. Giving feedback on interactive student programs with meta-exploration. *Advances in Neural Information Processing Systems*, 35:36282–36294, 2022.

[35] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

[36] Guy Lorberbom, Daniel D Johnson, Chris J Maddison, Daniel Tarlow, and Tamir Hazan. Learning generalized gumbel-max causal mechanisms. *Advances in Neural Information Processing Systems*, 34:26792–26803, 2021.

[37] Meili Lu, Xile Wei, Yanqiu Che, Jiang Wang, and Kenneth A Loparo. Application of reinforcement learning to deep brain stimulation in a computational model of parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(1):339–349, 2019.

[38] Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077, 2014.

[39] Ye Mao. One minute is enough: Early prediction of student success and event-level difficulty during novice programming tasks. In *In: Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*, 2019.

[40] Samiha Marwan, Anay Dombe, and Thomas W Price. Unproductive help-seeking in programming: What it is and how to address it. In *Proceedings of the 2020 ACM conference on innovation and technology in computer science education*, pages 54–60, 2020.

[41] Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. *Advances in Neural Information Processing Systems*, 31, 2018.

[42] Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S Thomas. Offline contextual bandits with high probability fairness guarantees. *Advances in neural information processing systems*, 32, 2019.

[43] Kohei Miyaguchi. A theoretical framework of almost hyperparameter-free hyperparameter selection methods for offline policy evaluation. *arXiv e-prints*, pages arXiv–2201, 2022.

[44] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in Neural Information Processing Systems*, 32, 2019.

[45] Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

[46] Allen Nie, Yannis Flet-Berliac, Deon Jordan, William Steenbergen, and Emma Brunskill. Data-efficient pipeline for offline reinforcement learning with limited data. *Advances in Neural Information Processing Systems*, 35:14810–14823, 2022.

[47] Allen Nie, Ann-Katrin Reuel, and Emma Brunskill. Understanding the impact of reinforcement learning personalization on subgroups of students in math tutoring. In *International Conference on Artificial Intelligence in Education*, pages 688–694. Springer, 2023.

[48] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

[49] Art B Owen. Monte carlo theory, methods and examples. 2013.

[50] Tom Le Paine, Cosmin Paduraru, Andrea Michi, Caglar Gulcehre, Konrad Zolna, Alexander Novikov, Ziyu Wang, and Nando de Freitas. Hyperparameter selection for offline reinforcement learning. *arXiv preprint arXiv:2007.09055*, 2020.

[51] Bahram Parvinian, Christopher Scully, Hanniebey Wiyor, Allison Kumar, and Sandy Weininger. Regulatory considerations for physiological closed-loop controlled medical devices used for automated critical care: food and drug administration workshop discussion topics. *Anesthesia and analgesia*, 126(6):1916, 2018.

[52] Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[53] Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.

[54] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4, pages 547–562. University of California Press, 1961.

[55] Florian Richter, Ryan K Orosco, and Michael C Yip. Open-sourced reinforcement learning environments for surgical robotics. *arXiv preprint arXiv:1903.02090*, 2019.

[56] Sherry Ruan, Allen Nie, William Steenbergen, Jiayu He, JQ Zhang, Meng Guo, Yao Liu, Kyle Dang Nguyen, Catherine Y Wang, Rui Ying, et al. Reinforcement learning tutor better supported lower performers in a math task. *arXiv preprint arXiv:2304.04933*, 2023.

[57] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

[58] Oleh Rybkin, Chuning Zhu, Anusha Nagabandi, Kostas Daniilidis, Igor Mordatch, and Sergey Levine. Model-based reinforcement learning via latent-space collocation. In *International Conference on Machine Learning*, pages 9190–9201. PMLR, 2021.

[59] Rolf Schwonke, Alexander Renkl, Carmen Krieg, Jörg Wittwer, Vincent Aleven, and Ron Salden. The worked-example effect: Not an artefact of lousy control conditions. *Computers in human behavior*, 25(2):258–266, 2009.

[60] Shitian Shen and Min Chi. Reinforcement learning: the sooner the better, or the later the better? In *Proceedings of the 2016 conference on user modeling adaptation and personalization*, pages 37–44, 2016.

[61] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[62] Robert R Sinclair, James E Martin, and Robert P Michel. Full-time and part-time subgroup differences in job attitudes and demographic characteristics. *Journal of Vocational Behavior*, 55(3):337–357, 1999.

[63] Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*, pages 9196–9205. PMLR, 2020.

[64] John Sweller and Graham A Cooper. The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and instruction*, 2(1):59–89, 1985.

[65] Shengpu Tang and Jenna Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pages 2–35. PMLR, 2021.

[66] 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.

[67] Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pages 9659–9668. PMLR, 2020.

[68] Kurt VanLehn. The behavior of tutoring systems. *International Journal Artificial Intelligence in Education*, 16(3):227–265, 2006.

[69] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[70] Cameron Voloshin, Nan Jiang, and Yisong Yue. Minimax model learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1612–1620. PMLR, 2021.

[71] Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

[72] Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[73] Pengcheng Wang, Jonathan Rowe, Wookhee Min, Bradford Mott, and James Lester. Interactive narrative personalization with deep reinforcement learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

[74] Peter R Wurman, Samuel Barrett, Kenta Kawamoto, James MacGlashan, Kaushik Subramanian, Thomas J Walsh, Roberto Capobianco, Alisa Devlic, Franziska Eckert, Florian Fuchs, et al. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, 2022.

[75] Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

[76] Mengjiao Yang, Bo Dai, Ofir Nachum, George Tucker, and Dale Schuurmans. Offline policy selection under uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pages 4376–4396. PMLR, 2022.

[77] Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. *Advances in Neural Information Processing Systems*, 33:6551–6561, 2020.

[78] Xi Yang, Yuan Zhang, and Min Chi. Multi-series time-aware sequence partitioning for disease progression modeling. In *IJCAI*, 2021.

[79] 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. *arXiv preprint arXiv:2104.13877*, 2021.

[80] Shangtong Zhang, Bo Liu, and Shimon Whiteson. Gradientdice: Rethinking generalized offline estimation of stationary values. In *International Conference on Machine Learning*, pages 11194–11203. PMLR, 2020.

[81] Siyuan Zhang and Nan Jiang. Towards hyperparameter-free policy selection for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12864–12875, 2021.

[82] Rujie Zhong, Duohan Zhang, Lukas Schäfer, Stefano Albrecht, and Josiah Hanna. Robust on-policy sampling for data-efficient policy evaluation in reinforcement learning. *Advances in Neural Information Processing Systems*, 35:37376–37388, 2022.

[83] Guojing Zhou, Hamoon Azizsoltani, Markel Sanz Ausin, Tiffany Barnes, and Min Chi. Hierarchical reinforcement learning for pedagogical policy induction. In *International conference on artificial intelligence in education*, pages 544–556. Springer, 2019.

[84] Guojing Zhou, Hamoon Azizsoltani, Markel Sanz Ausin, Tiffany Barnes, and Min Chi. Leveraging granularity: Hierarchical reinforcement learning for pedagogical policy induction. *International journal of artificial intelligence in education*, 32(2):454–500, 2022.

## List of Appendices

Appendix17

**A Detailed Setup of the IE Experiment and Additional Discussions** **17**
A.1 The IE System for the College Entry-Level Course. . . . . 17
A.2 Classroom Setup . . . . . 17
A.3 Environmental Setup of the IE System . . . . . 18
A.3.1 State Features. . . . . 18
A.3.2 Actions & rewards. . . . . 18
A.3.3 Behavior policy. . . . . 18
A.3.4 Target (evaluation) policies. . . . . 18
A.3.5 Sub-group identification. . . . . 19
A.4 Data Pre-Processing for Sub-Group Partitioning with the IE Experiment . . . . . 19
A.5 More Discussions over the Results from Section 3.2 . . . . . 20

**B Detailed Setup of the Healthcare Experiment** **23**
B.1 States & actions. . . . . 23
B.2 Rewards. . . . . 23
B.3 Optimal policy. . . . . 23
B.4 Behaivor policy. . . . . 23
B.5 Target policies. . . . . 23

**C Societal and Broader Impacts** **23**
C.1 Societal Impacts . . . . . 23
C.2 Broader Impact on Facilitating Fairness in RL-Empowered HCSs . . . . . 23

**D More on the Methodology** **24**
D.1 Practical Off-Policy Deployment in HCSs over the Sub-Group Objective (1). . . . 24
D.2 Proof of Bound 3 . . . . . 24
D.3 Detailed Formulation of VAE in MDP . . . . . 25
D.4 Proof of Equation 4 . . . . . 25

**E More Experimental Setup** **27**
E.1 Training Resources . . . . . 27
E.2 Implementations and Hyper-parameters . . . . . 27
E.3 OPE Standard Evaluation Metrics . . . . . 27

# A Detailed Setup of the IE Experiment and Additional Discussions

## A.1 The IE System for the College Entry-Level Course.

Though the problem setting and our method are general and can be applied to other interactive IE systems, we primarily focus on the system specifically used in an undergraduate probability course at a university, which has been extensively used by over $1,288$ students with $\sim$800k recorded interaction logs through 6 academic years. The IE system is designed to teach entry-level undergraduate students with ten major probability principles, including complement theorem, mutually exclusive theorem, independent events, De Morgan's theorem, addition theorem for two events, addition theorem for three events, conditional independent events, conditional probability, total probability theorem, and Bayes' rule.

Each students went through four phases, including (*i*) reading the textbook; (*ii*) pre-exam; (*iii*) studying on the IE system; and (*iv*) post-exam. During the reading textbook phase, students read a general description of each principle, review examples, and solve some training problems to get familiar with the IE system. Subsequently, they take a pre-exam comprising a total of 14 single- and multiple-principle problems. During the pre-exam, students are not provided with feedback on their answers, nor are they allowed to go back to earlier questions (so as the post-exam). Then, students proceed to work on the IE system, where they receive the same 12 problems in a predetermined order. After that, students take the 20-problem post-exam, where 14 of the problems are isomorphic to the pre-exam and the remainders are non-isomorphic multiple-principle problems. Exams are auto-graded following the same grading criteria set by course instructors.

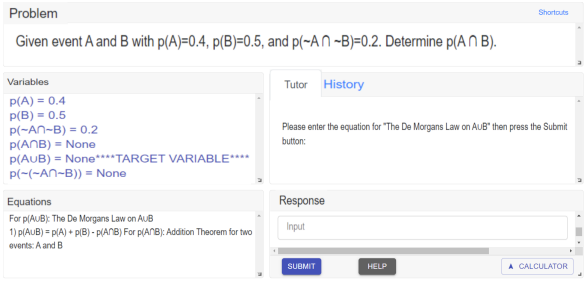



Figure 3: Graphical user interface (GUI) of the IE system. The problem statement window (*top*) presents the statement of the problem. The dialog window (*middle right*) shows the message the tutor provides to the students. Responses, e.g., writing an equation, are entered in the response window (*bottom right*). Any variables and equations generated through this process are shown on the variable window (*middle left*) and equation window (*bottom left*).

Since students' underlying characteristics and mind states are inherently unobservable [38], the IE system defined its state space with 142 features that could possibly capture students' learning status based on their interaction logs, as suggested by domain experts. While tutoring, the agent makes decisions on two levels of granularity: problem-level first and then step-level. For problem-level, it first decides whether the next problem should be a worked example (`WE`) [64], problem-solving (`PS`), or a collaborative problem-solving worked example (`CPS`) [59]. In `WE`s, students, observe how the tutor solves a problem; in `PS`s, students solve the problem themselves; in `CPS`s, the students and the tutor co-construct the solution. If a `CPS` is selected, the tutor will then make step-level decisions on whether to `elicit` the next step from the student or to `tell` the solution step to the student directly. Besides post-exam score, another important measure of student learning outcomes is their normalized learning gain (NLG), which is calculated by their pre- and post-exam scores $NLG = \frac{score_{postexam} - score_{preexam}}{\sqrt{1 - score_{preexam}}}$. The NLG defined in [5], represents the extent to which students have benefited from the IE system in terms of improving their learning outcomes.

## A.2 Classroom Setup

**Participants recruitment.** All participants were entry-level undergraduates majoring in STEM and enrolled in the Probability course in a college. They were recruited via invitation emails and told the procedure of the study and their data were used for research purpose only, and the study was an opt-in without influence on their course grades. Participants can also opt-in not recording their logs and quit the study any time. No demographics data or course grades were collected. All participants had acknowledged the study procedure and future research conducted using their logs.

**Principles taught by the IE system.** Table 3 shows all ten principles for the IE system to teach designed for the undergraduate entry-level students with STEM majors.

**Pre- and post-exams.** As introduced in Section 2, we use pre- and post-exams to measure the extent to which students have benefited from the IE system for improved learning outcomes. Tables 4 & 5 contain all problems in pre- and post-exams during our experiment with the IE system.

**The set of candidate target policies under consideration.** For safety concerns, only three RL-induced target policies that passed expert sanity checks can be deployed, while the expert policy still remained in the semester as the control group. For fairness concerns, the IE system randomly assigned a policy to each student, while we tracked the policies selected by FPS on each subgroup to evaluate its effectiveness. The chi-squared test was employed to check the relationship between policy assignment and subgroups, and it showed that the policy assignment cross subgroups were balanced with no significant relationship (p-value=0.479). In the testing semester, 140 students accomplished all problems and exams.

We provide the design of each problem regarding principles coverage for readers' interests. Detailed problem descriptions are omitted for identity and anonymity, which are only accessible within the research groups under IRB. An example problem description is shown in Figure 3.

### A.3 Environmental Setup of the IE System

#### A.3.1 State Features.

The state features were defined by domain experts that could possible capture students' learning status based on their interaction logs. In sum, 142 features with both discrete and continuous values are extracted, we provide summary descriptions of the features characterized by their systematic functions: (*i*) Autonomy (10 features): the amount of work done by the student, such as the number of times the student restarted a problem; (*ii*) Temporal Situation (29 features): the time-related information about the work process, such as average time per step; (*iii*) Problem-Solving (35 features): information about the current problem-solving context, such as problem difficulty; (*iv*) Performance (57 features): information about the student's performance during problem-solving, such as percentage of correct entries; (*v*) Hints (11 features): information about the student's hint usage, such as the total number of hints requested.

#### A.3.2 Actions & rewards.

See A.1 above.

#### A.3.3 Behavior policy.

The behavior policy follows an expert policy commonly used in e-learning [83], randomly taking the next problem as a worked example (`WE`), problem-solving by students (`PS`), or a collaborative problem-solving working examples (`CPS`). Note that the three decision choices are designed by domain experts that are found can support students' learning in prior works [59, 64], thus the expert policy is considered as effective.

#### A.3.4 Target (evaluation) policies.

In total, four target policies, including three RL-induced and the expert policy, were examined in testing semester. The three RL-induced policies were trained using off-policy DQN-based algorithm, and passed expert sanity check. In this study, expert sanity check were conducted by departments and independent instructors for pre-examination of the target policies.

Specifically, we employed the DQN-based algorithm designed by domain researchers [24], called Critical-RL, that have achieved empirical significance in real-world classrooms, and passed expert sanity check by our institutions. First, for each problem, a pair of adversarial policies using vanilla DQN algorithm were induced, including an original policy induced using the original rewards and an inversed policy induced using the inversed rewards (*i.e.*, the negative value of the original rewards). Then, critical decisions are identified following two rules: (1) Given the state, the two policies make opposite decisions; *and* (2) the decision is important (critical) for both policies. For a given state, rule (1) is tested first. If the adversarial policies make the same decisions, it is not critical. Otherwise,

rule (2) is tested. In order to measure the importance of the decision for each policy, we calculate the absolute Q-value difference between the two alternative actions. If this difference is greater than a threshold, the decision is considered important (critical). A critical-RL policy carries out the original policy when a decision is recognized as critical, and expert policy for the rest. Overall, in this study, we examined one critical-RL policy and two variations of it, *i.e.*, a policy carrying out original policy when a decision is *not* critical, and a policy carrying out original policy over all decisions. We set the threshold to be the median Q-value difference for all decisions in our training data set following the settings of the original Critical-RL work [24]. Each pair of of adversarial policies considered all parts of the training data were identical, such as state representation and transition samples, except the rewards. We use the learning rate $lr = 1e-3$ for inducing DQN policies.

#### A.3.5 Sub-group identification.

Specifically, to learn the subgroups in the IE system, we leverage an off-the-shelf algorithm called Toeplitz inverse covariance-based clustering (TICC) [21] to map initial logs $\mathcal{S}_0$ into $M$ clusters based on the values of student-sensitive features (as defined in Appendix A.4), where each $s_0 \in \mathcal{S}_0$ is associated with a cluster from the set $\mathbf{K} = \{K_1, \ldots, K_M\}$, over the offline dataset, from which the number of individuals within each cluster is intrinsically determined. Specifically, TICC initially clusters all states from the offline dataset, where the states that are mapped to the same cluster can be considered to share the graphical connectivity structure of cross-features and temporal information captured by TICC. Then the clusters of initial states can be determined accordingly from the clustering outcomes. We consider using TICC because of its superior performance in clustering compared to traditional distance-based methods such as K-means, especially with human behavior-related tasks [21, 78], such that the clusters of initial logs could be more scalable and capable of the evolving individuals and their behaviors in real-world HCSs. For a new participant arriving in the testing period, the cluster where the participant may belong to is the cluster exhibiting the least averaging distance between the initial states of the participant and samples within the cluster captured from the offline dataset.

The size of clusters is determined by a data-driven procedure following the original TICC work (*i.e.*, it is determined with the highest silhouette score in clustering historical trajectories) [21]. Note that we exhibit TICC as an example in our proposed pipeline, while it can be replaced by other partitioning approaches if needed. Then, we assume subgroup partitioning is consistent with cluster assignments associated with initial logs, *i.e.*, students whose initial logs are associated with the same cluster index are considered from the same subgroup.

**TICC Problem.** Each cluster $m \in [1, M]$ is defined as a Markov random field [57], or correlation network, captured by its Gaussian inverse covariance matrix $\Sigma_m^{-1} \in \mathbb{R}^{c \times c}$, where $c$ is the dimension of state space. We also define the set of clusters $\mathbf{K} = \{K_1, \ldots, K_M\} \subset \mathbb{R}$ as well as the set of inverse covariance matrices $\mathbf{\Sigma}^{-1} = \{\Sigma_1^{-1}, \ldots, \Sigma_M^{-1}\}$. Then the objective is set to be: $\max_{\mathbf{\Sigma}^{-1}, \mathbf{K}} \sum_{m=1}^{M} \left[ \sum_{s_t^{(i)} \in K_m} \left( \mathcal{L}(s_t^{(i)}; \Sigma_m^{-1}) - \epsilon \mathbb{1}\{s_{t-1}^{(i)} \notin K_m\} \right) \right]$, where the first term defines the log-likelihood of $s_t^{(i)}$ coming from $K_m$ as $\mathcal{L}(s_t^{(i)}; \Sigma_m^{-1}) = -\frac{1}{2}(s_t^{(i)} - \mu_m k)^T \Sigma_m^{-1} (s_t^{(i)} - \mu_m) + \frac{1}{2} \log \det \Sigma_m^{-1} - \frac{n}{2} \log(2\pi)$ with $\mu_m$ being the empirical mean of cluster $K_m$, the second term $\mathbb{1}\{s_{t-1}^{(i)} \notin K_m\}$ penalizes the adjacent events that are not assigned to the same cluster and $\epsilon$ is a constant balancing off the scale of the two terms. This optimization problem can be solved using the expectation-maximization family of algorithms by updating $\mathbf{\Sigma}^{-1}$ and $\mathbf{K}$ alternatively [21].

### A.4 Data Pre-Processing for Sub-Group Partitioning with the IE Experiment

The initial logs (serving as the initial states in the MDP) of students are used for sub-group partitioning, which is now only benefited from the FPS framework design but over two educational perspectives. First, initial logs may reflect not only the background knowledge of students but their interaction habits [13], without specific information related to behavior policies that may be distracting for sub-group partitioning. Though some existing works utilize demographics or grades of students from their prior taken courses to identify student subgroups [2, 62], it may not be feasible in practice due to the protection of student information by institutions. Second, prior works have found that initial logs can be informative to indicate learning outcomes of students [39], which makes it possible for the IE system to customize the policies with the goal of improving learning outcomes for each subgroup.

However, there is a challenge with sub-group partitioning over the initial logs of students. The state space of student logs in the IE system is usually high-dimensional, due to the detailed capture of each step taken during interaction and associated timing information [5, 38]. For example, in this study, 142 features have been recorded. While some features might be irrelevant for downstream data mining tasks, it is challenging to determine their relevance a priori [38]. To solve this, we used a data-driven feature taxonomy over the state features of students, then perform subgroup partitioning with distilled features based on the feature taxonomy.

**The data-driven feature taxonomy over state features of students.** Educational researchers have used feature taxonomy in qualitative ways to support instructors subgroup students and understand behaviors of students [40]. Unlike prior approaches that are expensive requiring much effort from human experts, we used a data-driven feature taxonomy for a straightforward student subgroup partitioning that may reflect the knowledge background and dynamic learning progress of students. Specifically, we define three major categories of features according to their temporal and cross-student characteristics: (i) *System-assigned*: the features, which are static across students on the same problem, are assumed to be assigned by the system; (ii) *Student-centric*: the features, which differ across students from the initial logs and may change over time, is assumed to be students-centric and reflect both students' initial knowledge background and the changes of individual underlying mindset during learning; (iii) *Interaction-driven*: the features, which contain characteristics from both system-assigned and student-centric types, are assumed to be mixed-style features that are affected by both system and individuals. For example, the number of `tells` since `elicit` is set with a default value by the system while changing over time depending on students' progress.

**Sub-group partitioning with distilled features via feature taxonomy.** Since system-assigned features are mainly dominated by system design and remain static across students on each problem, for the purpose of sub-group partitioning, we focus on the two types of features, student-centric and interaction-driven, since both could be highly associated with students' underlying mental status and behaviors, for which we call *student-sensitive* features. In this work, we identified $82\%(117)$ from overall 142 features as student-sensitive features and used them for subgroup partitioning.

Table 2: Feature taxonomy with examples and percentage in the the IE system.

| Taxonomy | Examples | Perc. |
|---|---|---|
| System-assigned | Problem difficulty | 18% |
| Student-centric | Number of hints requested | 48% |
| Interaction-driven | Number of `tells` since `elicit` | 34% |

### A.5 More Discussions over the Results from Section 3.2

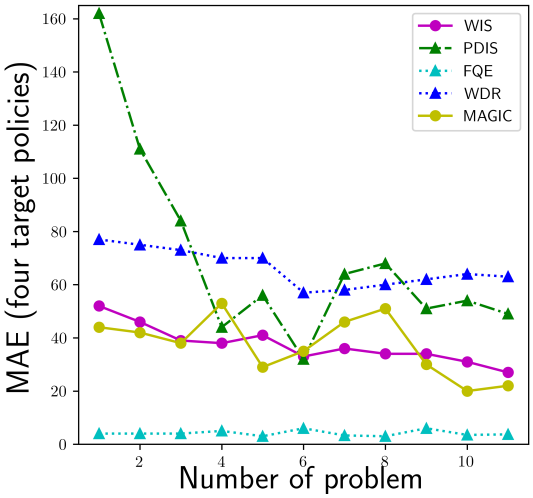



Figure 4: Mean absolute error (MAE) of OPE AugRRS with subgroup partitioning over problems in historical data.

**Would sub-group partitioning over a longer trajectory improve the performance of the OPS+VRRS baselines?** Recall that OPS+VRRS deployed the sub-optimal to most students, while

their estimation accuracy (*i.e.,* absolute error) was improved compared to purely OPS and OPS+RRS (Figure 1(b)), which is outperformed by FPS over a slight margin. We further investigate the effects of subg-roup partitioning with longer trajectory information on OPS+VRRS performance. We conduct sub-group partitioning over the length of trajectories, *i.e.*, perform sub-group partitioning on the averaged states' features associated with the first $\Delta$ problems across historical trajectories, where $\Delta \in [1, 11]$ excluding the final problem. Then we augment the same amount of samples for each subgroup $K$, *i.e.*, $|\widehat{\mathcal{T}}_K| = |\mathcal{T}_K| = |K|$ and perform OPS+RRS. We observe that in all 55 conditions except the five (*i.e.*, WIS+VRRS $\Delta = 4, 11$, PDIS+VRRS $\Delta = 8$, and FQE+VRRS $\Delta = 7, 8$), all OPS+VRRS still select the sub-optimal policy. Figure 4 presents the mean absolute error (MAE) of the OPS+VRRS methods over the four target policies. It shows the trend of improved MAE over the number of problems for most methods. Those indicate that more information over a longer trajectory does have some positive effects on OPS+VRRS, but their policy selection is hard to be improved and stabilized. More students-centric and robust OPS methods are needed for IE policy selection.

Table 3: Principles taught by the IE system for undergraduate entry-level students.

| Abbr. | Name of principle | Expression |
|---|---|---|
| CT | Complement Theorem | $P(A) + P(\neg A) = 1$ |
| MET | Mutually Exclusive Theorem | $P(A \cap B) = 0$ iff A and B are mutually exclusive events |
| IE | Independent Events | $P(A \cap B) = P(A)P(B)$ if A and B are independent events |
| DMT | De Morgan's Theorem | $P(\neg(A \cup B)) = P(\neg A \cap \neg B), P(\neg(A \cap B)) = P(\neg A \cup \neg B)$ |
| A2 | Addition Theorem for two events | $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ |
| A3 | Addition Theorem for three events | $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - p(B \cap C) + P(A \cap B \cap C)$ |
| CIE | Conditional Independent Events | $P(A \cap B\|C) = P(A\|C)P(B\|C)$ if A and B are independent events given C |
| CP | Conditional Probability | $P(A \cap B) = P(A\|B)P(B) = P(B\|A)P(A)$ |
| TPT | Total Probability Theorem | $P(A) = P(A\|B_1)P(B_1) + P(A\|B_2)P(B_2) + \ldots + P(A\|B_n)P(B_n)$ if $B_1, B_2, \cdots, B_n$ are mutually exclusive events and $B_1 \cup B_2 \cup \cdots \cup B_n = W$ |
| BR | Bayes Rule | $P(B_i\|A) = P(A\|B_i)P(B_i)/P(A\|B_1)P(B_1) + P(A\|B_2)P(B_2) + \cdots + P(A\|B_n)P(B_n)$ if $B_1, B_2, \cdots, B_n$ are mutually exclusive events and $B_1 \cup B_2 \cup \cdots \cup B_n = W$ |

Table 4: Pre-exam problems in the IE system.

| Problem | CT | MET | IE | DMT | A2 | A3 | CIE | CP | TPT | BR |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | ✓ | | | | |
| 2 | | | | | | | | | ✓ | |
| 3 | | | | | | | | ✓ | | |
| 4 | ✓ | | | ✓ | ✓ | | | | | |
| 5 | | | | | | | | | | ✓ |
| 6 | | | | | | | ✓ | | | ✓ |
| 7 | | ✓ | ✓ | | | ✓ | | | | |
| 8 | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | |
| 9 | ✓ | | | | | | | | | |
| 10 | | | | | ✓ | | | | | |
| 11 | | | ✓ | | | | | | | |
| 12 | | | | ✓ | | | | | | |
| 13 | | | | | | | ✓ | | | |
| 14 | | ✓ | | | | | | | | |

Table 5: Post-exam problems in the IE system.

| Problem | CT | MET | IE | DMT | A2 | A3 | CIE | CP | TPT | BR | Iso-Test-Problem |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | ✓ | | | | | | | | 11 |
| 2 | | ✓ | ✓ | | | ✓ | | | | | 7 |
| 3 | ✓ | | | ✓ | ✓ | | | | | | 4 |
| 4 | ✓ | | | | | | | | | | 9 |
| 5 | | | | | | | | ✓ | | | 3 |
| 6 | | | | | ✓ | | | | | | 10 |
| 7 | | | | | | | | | ✓ | | 2 |
| 8 | | | | | | | ✓ | | | | 13 |
| 9 | | | | | | | | ✓ | ✓ | ✓ | N/A |
| 10 | | ✓ | | | | | | | | | 14 |
| 11 | | | | | | | | | | ✓ | 5 |
| 12 | | | | ✓ | | | | | | | 12 |
| 13 | | | | | | | ✓ | | | ✓ | 6 |
| 14 | ✓ | | ✓ | | | | | | | | N/A |
| 15 | ✓ | | | ✓ | ✓ | | | ✓ | | | N/A |
| 16 | ✓ | | | | | | | | | ✓ | N/A |
| 17 | | | | | | ✓ | | | | | 1 |
| 18 | ✓ | ✓ | | ✓ | ✓ | ✓ | | | | | 8 |
| 19 | ✓ | ✓ | | | ✓ | ✓ | | | | | N/A |
| 20 | | ✓ | | | | ✓ | | | ✓ | | N/A |

# B Detailed Setup of the Healthcare Experiment

We use the sepsis simulator developed by [48] and benchmark settings of [45].

### B.1 States & actions.

The definition of states and actions are introduced in Section 3.3.

### B.2 Rewards.

We also follow the benchmark [45] in terms of configuring the reward function and behavioral policy. Specifically, a reward of -1 is received at the end of horizon ($T = 5$) if the patient is deceased (*i.e.*, at least three vitals are out of the normal range), or +1 if discharged (when all vital signs are in the normal range without treatment).

### B.3 Optimal policy.

To learn the optimal policy, [45] used policy iteration to learn the optimal policy, and created a near-optimal (soft optimal) policy by having the policy take a random action with probability 0.05, and the optimal action with probability 0.95. The value function (for the optimal policy) was computed using value iteration. The discount factor $\gamma = 0.99$.

### B.4 Behaivor policy.

The behaviour policy is a mixture of two policies: 85% the soft optimal policy and 15% of a sub-optimal policy that is similar to the soft optimal but the vasopressors action is flipped.

### B.5 Target policies.

See Section 3.3.

# C Societal and Broader Impacts

### C.1 Societal Impacts

All real-world data employed in this paper were obtained anonymously through exempt IRB-approved protocols and were scored using established rubrics. No demographic data or class grades were collected. All data were shared within the research group under IRB, and were de-identified and automatically processed for labeling. This research seeks to remove societal harms that come from lower engagement and retention of students who need more personalized interventions and developing more robust medical interventions for patients.

### C.2 Broader Impact on Facilitating Fairness in RL-Empowered HCSs

Fairness in AI-empowered HCSs has been a long-standing concern [33, 42, 47, 56, 8]. The FPS framework can be potentially extended to promote fairness to a certain extent, by helping minority/under-represented groups to boost their utility gain through deployment of customized policy specific to the group. Specifically, following FPS, the sub-group partitioning step can identify the small-scaled yet important groups, then the policy that is most beneficial for the group can be deployed to maximize the gain. As illustrated by Figure 2, FPS effectively identifies the low-performing students (group $K_4$), which only constitute $< 5\%$ of the overall population at the 6-th (testing) semester, without leveraging any subjective information *a priori* (i.e., sub-group partitioning uses exactly the same features across all students). FPS then significantly boosts their performance by deploying the policy most suitable for the group. Similarly, one can easily extend the FPS framework to intelligent HCSs oriented toward other applications, in order to identify the groups that potentially need more attention, and help all participants to achieve similar gain indiscriminately by deploying the right policy to each participant.

# D More on the Methodology

## D.1 Practical Off-Policy Deployment in HCSs over the Sub-Group Objective (1).

The overall off-policy deployment can be achieved using a two-step approach, *i.e.*, $(i)$ *pre-partitioning* with offline dataset, followed by $(ii)$ *deployment* upon observation of the initial states of arriving participants.

To facilitate $(i)$, clustering methods, such as TICC [21], is used to learn a preliminary partitioning $l : \mathcal{S}_0 \to \mathcal{K}$. Then, the value function estimator $\hat{D}^{\pi,\beta}_{K_m=l(s_0)}$ is trained for estimating $V^{\pi}_{K_m=l(s_0)} - V^{\beta}_{K_m=l(s_0)}$ using part of the offline trajectories whose initial state $s_0$ falls in the corresponding group $K_m = l(s_0)$ following Proposition 2.5, for all $K_m \in \mathcal{K}$.[7] At last, one can learn a mapping $d : \mathcal{S}_0 \times \Pi \times \mathcal{K} \to \mathbb{R}$ to reconstruct $D^{\pi,\beta}_{K_m}$'s estimation using the $(s_0, \pi, K_m)$ triplets as the inputs (*e.g.*, by minimizing squared error).

In step $(ii)$, plug into $d(s_0^{(i)}, \pi, K_m)$ all policy candidates $\pi \in \Pi$ and all possible partitions $K_m \in \mathcal{K}$. Then, one can determine which policy $\pi$ satisfies $\max_{\pi\in\Pi} d(s_0^{(i)}, \pi, K_m)$. Assuming the pre-partitioning over offline dataset is knowledgeable about initial logs of participants, for each *arriving* participant $i' \geq N$ with their initial log $s_0^{(i')}$, determine the cluster they may most likely belong to. For example, the cluster $K_m$ with the least averaging distance between $s_0^{(i')}$ and $s_0^{(i)}$, for every $s_0^{(i)} \in K_m$. Then, assign to participant $i'$ the corresponding group $k(s_0^{(i')}) = K_m$, as captured by the mapping function $k : \mathcal{S}_0 \to \mathcal{K}$. Deploy the corresponding policy $\pi$ that is estimated as achieving best performance for group $K_m$, to $i'$.

## D.2 Proof of Bound 3

**Rényi divergence.** For $\epsilon \geq 0$, the Rényi divergence for two distribution $\pi$ and $\beta$ is defined by [54]

$$d_\epsilon(\pi\|\beta) = \frac{1}{\epsilon - 1} log_2 \sum_x \beta(x) \Big(\frac{\pi(x)}{\beta(x)}\Big)^{\epsilon-1}.$$

[6] denote the exponential in base 2 by $d_\epsilon(\pi\|\beta) = 2^{D_\epsilon(\pi\|\beta)}$.

*Proof.* $\hat{D}^{\pi,\beta}_{K_m} = \frac{1}{|\mathcal{I}_m|}\sum_{i\in\mathcal{I}_m}\left(\omega_i \sum_{t=1}^T \gamma^{t-1} r_t^{(i)} - \sum_{t=1}^T \gamma^{t-1} r_t^{(i)}\right)$, with $\omega_i = \Pi_{t=1}^T \pi(a_t^{(i)}|s_t^{(i)})/\beta(a_t^{(i)}|s_t^{(i)})$, can be upper bounded by the variance of importance sampling weights. Denote $g_i = \sum_{t=1}^T \gamma^{t-1} r_t^{(i)}$. Following [25], since $Var(\hat{G}) \leq \mathbb{E}[\hat{G}^2]$ (following the definition of variance),

$$\begin{aligned} Var(\hat{D}^{\pi,\beta}_{K_m}) &\leq \frac{\|g\|^2_\infty}{N^2}\mathbb{E}\big[\sum_i (\omega_i - 1)^2\big] \\ &= \frac{1}{N}\|g\|^2_\infty Var(\omega), \end{aligned}$$

with the last equality following the fact $\mathbb{E}[\omega] = 1$. Moreover, $Var(w) = d_2(\pi\|\beta) - 1$ as pointed out by [6], following the Rényi divergence [54]. Thus, the variance of the estimator $Var(\hat{D}^{\pi,\beta}_{K_m})$ is:

$$\begin{aligned} Var(\hat{D}^{\pi,\beta}_{K_m}) &\leq \|g\|^2_\infty \big(\frac{d_2(\pi\|\beta) - 1}{N}\big) \\ &= \|g\|^2_\infty \big(\frac{d_2(\pi\|\beta)}{N} - \frac{1}{N}\big). \end{aligned}$$

The expression can be related to the effective sample size (ESS) of the original data given the target policy [41], resulting in

$$Var(\hat{D}^{\pi,\beta}_{K_m}) \leq \|g\|^2_\infty \big(\frac{1}{ESS} - \frac{1}{N}\big),$$

which completes the proof. □

[7] Note that $\hat{D}^{\pi,\beta}_{K_m=l(s_0)}$ essentially approximates the sum of value difference in (1) over each $\mathcal{I}_m$.

**Remark.** Note that in the special case that behavior policy being the same as target policy, the bound evaluates to zero. Moreover, as noted by [25], denote the right-hand side of inequality 3 by $Var_u(\cdot)$, it can be used in each sub-group as a proxy of variance of the estimator in the subgroup, i.e.,

$$Var_u(\hat{D}^{\pi,\beta}_{K_m}) = \|g_m\|^2_\infty \big(\frac{1}{ESS(K_m)} - \frac{1}{|K_m|}\big);$$

here, $g_m$ refers to the total return of the trajectories pertaining to the sub-group $K_m$, and $ESS(K_m)$ can be estimated by $ESS$ using $\widehat{ESS}(K_m) = \frac{(\sum_{i\in\mathcal{I}_m} g_i)^2}{\sum_{i\in\mathcal{I}_m} g_i^2}$ [49].

### D.3 Detailed Formulation of VAE in MDP

**The latent prior** $p(z_0) \sim \mathcal{N}(0, I)$ representing the distribution of the initial latent states (at the beginning of each PST in the set $\mathcal{T}^g$), where $I$ is the identity covariance matrix.

**Encoder.** $q_\eta(z_t|s_{t-1}, a_{t-1}, s_t)$ is used to approximate the posterior distribution $p_\xi(z_t|s_{t-1}, a_{t-1}, s_t) = \frac{p_\xi(z_{t-1}, a_{t-1}, z_t, s_t)}{\int_{z_t\in\mathcal{Z}} p(z_{t-1}, a_{t-1}, z_t, s_t)dz_t}$, where $\mathcal{Z} \subset \mathbb{R}^m$ and $m$ is the dimension. Given that $q_\eta(z_{0:T}|s_{0:T}, a_{0:T-1}) = q_\eta(z_0|s_0)\prod_{t=1}^T q_\eta(z_t|z_{t-1}, a_{t-1}, s_t)$, both distributions $q_\eta(z_0|s_0)$ and $q_\eta(z_t|z_{t-1}, a_{t-1}, s_t)$ follow diagonal Gaussian, where mean and diagonal covariance are determined by multi-layer perceptrons (MLPs) and long short-term memory (LSTM), with neural network weights $\eta$. Thus, one can infer $z_0^\eta \sim q_\eta(z_0|s_0)$, $z_t^\eta \sim q_\eta(z_t|h_t^\eta)$, with $h_t^\eta = f_\eta(h_{t-1}^\eta, z_{t-1}^\eta, a_{t-1}, s_t)$ where $f_\eta$ represents LSTM layer and $h_t^\eta$ represents LSTM recurrent hidden state.

**Decoder.** $p_\xi(z_t, s_t, r_{t-1}|z_{t-1}, a_{t-1})$ is used to sample new trajectories. Given $p_\xi(z_{1:T}, s_{0:T}, r_{0:T-1}|z_0, \xi) = \prod_{t=0}^T p_\xi(s_t|z_t)\prod_{t=1}^T p_\xi(z_t|z_{t-1}, a_{t-1})p_\xi(r_{t-1}|z_t)$, where $a_t$'s are determined following the behavioral policy $\beta$, distributions $p_\xi(s_t|z_t)$ and $p_\xi(r_{t-1}|z_t)$ follow diagonal Gaussian with mean and covariance determined by MLPs and $p_\xi(z_t|z_{t-1}, a_{t-1})$ follows diagonal Gaussian with mean and covariance determined by LSTM.

Thus, the generative process can be formulated as, *i.e.*, at initialization, $z_0^\xi \sim p(z_0)$, $s_0^\xi \sim p_\xi(s_0|z_0^\xi)$, $a_0 \sim \beta(a_0|s_0^\xi)$; followed by $z_t^\xi \sim p_\xi(\tilde{h}_t^\xi)$, $r_{t-1}^\xi \sim p_\xi(r_{t-1}|z_t^\xi)$, $s_t^\xi \sim p_\xi(s_t|z_t^\xi)$, $a_t \sim \beta(a_t|s_t^\xi)$, with $\tilde{h}_t^\xi = g_\xi[f_\xi(h_{t-1}^\xi, z_{t-1}^\xi, a_{t-1})]$ where $g_\xi$ represents an MLP.

**Training objective.** The training objective for the VAE in MDP is to maximize the evidence lower bound (ELBO), which consists of the log-likelihood of reconstructing the states and rewards, and regularization of the approximated posterior, *i.e.*,

$$\begin{aligned} ELBO(\eta,\xi) = \mathbb{E}_{q_\eta}\Big[&\sum\nolimits_{t=0}^T \log p_\xi(s_t|z_t) + \sum\nolimits_{t=1}^T \log p_\xi(r_{t-1}|z_t) \\ &- KL\big(q_\eta(z_0|s_0)||p(z_0)\big) - \sum\nolimits_{t=1}^T KL\big(q_\eta(z_t|z_{t-1}, a_{t-1}, s_t)||p_\xi(z_t|z_{t-1}, a_{t-1})\big)\Big]. \end{aligned} \tag{4}$$

The proof of Equation 4 is provided in Appendix D.4.

**More discussions on trajectory augmentations.** Latent-model-based models such as VAE have been commonly used for augmentation in offline RL, while they general rarely come with error bounds provided [20, 31, 58]. Prior works have also found that applying generative models to data augmentation can learn more robust predictors that are invariant especially with subgroup identity [18]. Though generative augmentation models may not perfectly model the subgroup distribution and introduce artifacts, as noted by [18], we can directly control the deviations of augmentation from original data with translation or consistency loss as in Equation 4. Our experimental results further show that off-policy selection can benefit more with combination of augmented samples and raw data compared to using raw (original) data only.

### D.4 Proof of Equation 4

The derivation of the evidence lower bound (ELBO) for the joint log-likelihood distribution can be found below.

*Proof.*

$$
\begin{aligned}
& \log p_\beta(s_{0:T}, r_{0:T-1}) && (5)\\
=& \log \int_{z_{1:T}\in\mathcal{Z}} p_\beta(s_{0:T}, z_{1:T}, r_{0:T-1}) dz && (6)\\
=& \log \int_{z_{1:T}\in\mathcal{Z}} \frac{p_\beta(s_{0:T}, z_{1:T}, r_{0:T-1})}{q_\alpha(z_{0:T}|s_{0:T}, a_{0:T-1})} q_\alpha(z_{0:T}|s_{0:T}, a_{0:T-1}) dz && (7)\\
\overset{Jensen's\ inequality}{\geq}& \mathbb{E}_{q_\alpha}[\log p(z_0) + \log p_\beta(s_{0:T}, z_{1:T}, r_{0:T-1}|z_0) - \log q_\alpha(z_{0:T}|s_{0:T}, a_{0:T-1})] && (8)\\
=& \mathbb{E}_{q_\alpha}\Big[\log p(z_0) + \log p_\beta(s_0|z_0) + \sum\nolimits_{t=0}^{T} \log p_\beta(s_t, z_t, r_{t-1}|z_{t-1}, a_{t-1})\\
& \qquad - \log q_\alpha(z_0|s_0) - \sum\nolimits_{t=1}^{T} \log q_\alpha(z_t|z_{t-1}, a_{t-1}, s_t)\Big] && (9)\\
=& \mathbb{E}_{q_\alpha}\Big[\log p(z_0) - \log q_\alpha(z_0|s_0) + \log p_\beta(s_0|z_0) + \sum\nolimits_{t=1}^{T} \log\big(p_\beta(s_t|z_t) p_\beta(r_{t-1}|z_t) p_\beta(z_t|z_{t-1}, a_{t-1})\big)\\
& \qquad - \sum\nolimits_{t=1}^{T} \log q_\alpha(z_t|z_{t-1}, a_{t-1}, s_t)\Big] && (10)\\
=& \mathbb{E}_{q_\alpha}\Big[\sum\nolimits_{t=0}^{T} \log p_\beta(s_t|z_t) + \sum\nolimits_{t=1}^{T} \log p_\beta(r_{t-1}|z_t)\\
& \qquad - KL\big(q_\alpha(z_0|s_0)||p(z_0)\big) - \sum\nolimits_{t=1}^{T} KL\big(q_\alpha(z_t|z_{t-1}, a_{t-1}, s_t)||p_\beta(z_t|z_{t-1}, a_{t-1})\big)\Big]. && (11)
\end{aligned}
$$

□

# E More Experimental Setup

## E.1 Training Resources

All experimental workloads are distributed among 4 Nvidia RTX A5000 24GB, 3 Nvidia Quadro RTX 6000 24GB, and 4 NVIDIA TITAN Xp 12GB graphics cards.

## E.2 Implementations and Hyper-parameters

For FQE, as in [30], we train a neural network to estimate the values of the target polices $\pi \in \mathbf{\Pi}$ by bootstrapping from the learned Q-function. For DualDICE, we use the open-sourced code in its original paper. For MAGIC, we use the implementation of [71]. For trajectory augmentation, for the components involving LSTMs, which include $q_\alpha(z_t|z_{t-1}, a_{t-1}, s_t)$ and $p_\beta(z_t|z_{t-1}, a_{t-1})$, their architecture include one LSTM layer with 64 nodes, followed by a dense layer with 64 nodes. All other components do not have LSTM layers involved, so they are constituted by a neural network with 2 dense layers, with 128 and 64 nodes respectively. The output layers that determine the mean and diagonal covariance of diagonal Gaussian distributions use linear and softplus activations, respectively. The ones that determine the mean of Bernoulli distributions (*e.g.*, for capturing early termination of episodes) are configured to use sigmoid activations. For training, in subgroups with sample size greater than 200, the maximum number of iteration is set to 1000 and minibatch size set to 64, and 200 and 4 respectively for subgroups with sample size less than or equal to 200. Adam optimizer is used to perform gradient descent. To determine the learning rate, we perform grid search among $\{1e-4, 3e-3, 3e-4, 5e-4, 7e-4\}$. Exponential decay is applied to the learning rate, which decays the learning rate by 0.997 every iteration. For OPE, the model-based methods are evaluated by directly interacting with each target policy for 50 episodes, and the mean of discounted total returns ($\gamma = 0.9$) over all episodes is used as estimated performance for the policy.

## E.3 OPE Standard Evaluation Metrics

*Absolute error* The absolute error is defined as the difference between the actual value and the estimated value of a policy:

$$AE = |V^\pi - \hat{V}^\pi| \tag{12}$$

where $V^\pi$ represents the actual value of the policy $\pi$, and $\hat{V}^\pi$ represents the estimated value of $\pi$.

*Mean absolute error (MAE)* The MAE is defined as the average value of absolute error across $|\mathbf{\Pi}|$ target (evaluation) policies:

$$MAE = \frac{1}{|\mathbf{\Pi}|} \sum_{\pi \in \mathbf{\Pi}} AE(\pi). \tag{13}$$

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: To the best of our knowledge, we are the first work that targeted and solved a practical challenge often encountered in off-policy selection (OPS) when deploying RL policies to new human arrivals in practical human-centric systems (HCSs). We have provided an algorithm and theoretical analysis to address the practical problem of interest, with extensive empirical justifications as provided by two human-centric environments across two domains (education and healthcare).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Please see Section 2 and Appendix D.2.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Please see Section 3 and Appendices A, B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Please see Supplementary Materials. Real-world human data is not publicly released under IRB protocols.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see Appendices A, B, E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please see Figures 1, 2 and Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see Appendix C.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper has not been found to pose no such risks yet.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The authors respectfully cited the original paper that produced the sepsis dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The authors provided the details of the code as part of their submissions via structured templates.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Please see Appendices A, C.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Please see Appendix C. All real-world data employed in this paper were obtained anonymously through exempt IRB-approved protocols and were scored using established rubrics. No demographic data or class grades were collected. All data were shared within the research group under IRB, and were de-identified and automatically processed for labeling. This research seeks to remove societal harms that come from lower engagement and retention of students who need more personalized interventions and developing more robust medical interventions for patients.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.