Reinforcement Learning for Clinical Site-of-Care Triage in a Sepsis Simulator

Extended Abstract

Liane Ozoemelam Saimai Lau Yun Dong
{lianeozo, smlau, yund2025}@stanford.edu

Motivation. Clinical triage is the decision of *where* a patient should be treated — the **site of care** (SOC), ranging from asynchronous monitoring through ambulatory and facility care to the ICU — and *how* they should be treated. These decisions are coupled, sequential, and made under partial observability of the patient's true physiological state. ICU capacity is scarce and costly, so both over-triage and under-triage carry real harm. We ask a focused question: *can reinforcement learning (RL) learn acuity-appropriate site-of-care triage in a sepsis simulator, and what governs whether it succeeds — the algorithm, the reward, or the simulator?*

Method. We extend the Gumbel-Max structural causal model (SCM) sepsis simulator of Oberst & Sontag [10] into a triage-focused partially observable Markov decision process (POMDP) in which SOC is both a state variable and an action dimension. The agent chooses one of $32 = 4_{\text{SOC}} \times 2^3_{\text{treatment}}$ joint actions subject to per-SOC feasibility constraints, observes only a SOC-masked subset of the true state, and is trained against five reward variants that progressively reshape terminal, treatment, and resource costs. We benchmark a rule-based heuristic, seven online RL algorithms (DQN, Double DQN, PPO, FactPPO, SAC, and two SAC variants), and two offline algorithms (IQL, IQL-KL-F), plus an exploratory model-based planner, all sharing a feasibility-aware design.

Implementation. We built a shared training and evaluation harness supporting on-policy, off-policy, offline, and model-based learning, trained every algorithm across all five reward variants with five random seeds, and evaluated on fixed patient pools using clinical and behavioral metrics (mortality, discharge, infeasible-action rate, and site-of-care usage). Offline algorithms were trained on a $\approx$0.9M-transition mixed dataset assembled from the online policies' trajectories.

Results. The best method is offline **IQL-KL-F**, reaching 13.6% mortality and 42.4% discharge with 0% infeasible actions. Strikingly, the *original* reward (reward0) ranks best for every algorithm except DQN: reward shaping changes behavior but does not consistently improve outcomes. Offline methods beat online ones; actor-critic methods beat value-based ones; and a simple feasibility penalty drives infeasible-action rates to zero.

Discussion. Because neither algorithm choice nor reward design recovers clinically reasonable cross-site-of-care triage, the evidence points to a mismatch between the ICU-derived, partly hand-specified simulator dynamics and the broader SOC triage task. This reframes the central obstacle from "which algorithm/reward" to "which environment is appropriate for cross-site-of-care evaluation."

Conclusion. We deliver a controlled, multi-algorithm, multi-reward study of RL triage and a clear negative result that localizes the limiting factor to the environment-task mismatch. Our intended next step is to re-ground and validate the simulator's transition dynamics against real clinical data; this is currently gated only by data-access credentialing.

Reinforcement Learning for Clinical Site-of-Care Triage in a Sepsis Simulator

Liane Ozoemelam
Stanford University
lianeozo@stanford.edu

Saimai Lau
Stanford University
smlau@stanford.edu

Yun Dong
Stanford University
yund2025@stanford.edu

Abstract

We study whether reinforcement learning can learn acuity-appropriate site-of-care triage for sepsis patients in a simulated clinical environment, and we isolate what limits success. We extend the Gumbel-Max SCM sepsis simulator into a triage POMDP in which the site of care is simultaneously part of the state and a dimension of the action, yielding a 32-way joint action space with per-site feasibility constraints and SOC-dependent partial observability. Across ten algorithms (a rule-based heuristic, seven online RL methods, and two offline methods) plus an exploratory model-based planner, and five reward variants, we find that the best policy is offline IQL-KL-F (13.6% mortality, 42.4% discharge, 0% infeasible actions), that offline methods dominate online ones, and that a lightweight feasibility regularizer eliminates infeasible actions. Our central finding is a negative one: the original reward is best for nearly every algorithm, so reward shaping is not the lever; instead, the limiting factor appears to be a mismatch between the MIMIC-III-derived, partly hand-specified simulator dynamics and the broader cross-site-of-care triage task. We argue this motivates re-grounding and validating the environment model against real clinical data, and we release a reusable harness for such studies.

1 Introduction

The rapid expansion of digital health technologies has multiplied the ways care can be delivered, and with it the complexity of deciding *where* and *how* to treat a patient. For an acute, time-sensitive condition such as sepsis, these site-of-care (SOC) and treatment decisions are tightly coupled and must be made sequentially under uncertainty about the patient's true physiological state. Well-designed decision-support tools could reduce clinician burden and improve allocation of scarce resources such as ICU beds, where both over-triage (wasting capacity) and under-triage (delaying critical care) are costly.

Reinforcement learning (RL) is a natural framework for such sequential decision-making, but clinical deployment is bounded by the quality of the environment in which policies are learned and evaluated. In this project we ask a deliberately scoped question: *can RL learn acuity-appropriate triage in a sepsis simulator, and what governs whether it succeeds?* Rather than chase a single best policy, we treat the study as a controlled diagnostic that varies the *algorithm*, the *reward*, and (by implication) the *simulator*, so as to localize where the difficulty actually lies.

Our contributions are:

1. A triage-focused POMDP extension of the Gumbel-Max SCM sepsis simulator [10] that promotes site of care to both a state variable and an action dimension, with a 32-way joint action space, per-SOC treatment feasibility, and SOC-dependent partial observability.

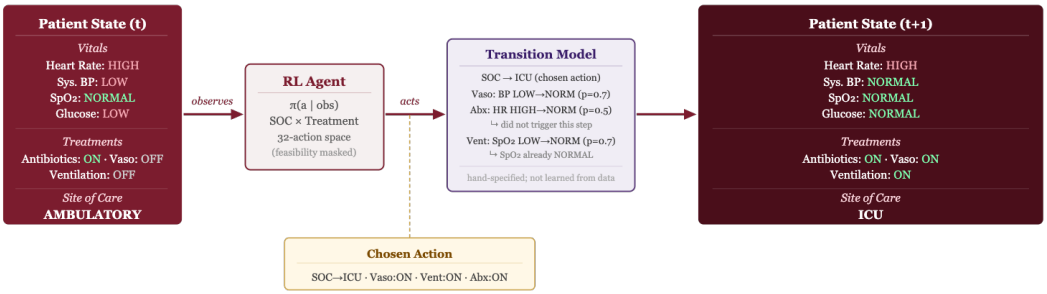

Figure 1: A sample triage transition. The agent observes a SOC-masked subset of the true patient state, selects a joint (site-of-care $\times$ treatment) action from a 32-way space filtered for per-SOC feasibility, and the hand-specified stochastic transition model advances the patient. Reward is computed from the state change (see Section 3.2).

2. A broad, controlled benchmark: ten algorithms (a heuristic, seven online RL methods, and two offline methods), plus an exploratory model-based planner, crossed with five reward variants, evaluated under a common harness with clinical and behavioral metrics.
3. A lightweight, transferable feasibility regularizer that drives the infeasible-action rate to zero across both online (SAC) and offline (IQL) methods.
4. A clear negative result: the original reward is best for nearly every algorithm, which — together with the dominance of offline RL and the failure of reward shaping — suggests that the limiting factor is a mismatch between the MIMIC-III-derived, partly hand-specified simulator dynamics and the broader cross-site-of-care triage task, rather than algorithm or reward choice alone.

2 Related Work

Simulated sepsis environments. Oberst & Sontag [10] introduced a Gumbel-Max SCM sepsis simulator whose discrete transition dynamics were learned from the MIMIC-III critical-care database [5], but their goal was counterfactual off-policy *evaluation*, not the design of triage policies. A prior Stanford course project, "From Acuity to Allocation" [1], added site-of-care structure to this simulator but reported persistent reward misalignment and degenerate policies. Our work builds directly on this lineage, but reframes the problem around an explicit triage action space and a systematic algorithm$\times$reward sweep designed to attribute failure.

Value-based and policy-gradient RL. We include DQN [9] and Double DQN [13], which reduces the maximization bias of Q-learning by decoupling action selection from evaluation, as well as PPO [11] and (discrete) SAC [3, 2]. **Offline RL.** Because online interaction in clinical settings is infeasible, offline RL [8] is especially relevant; we use Implicit Q-Learning (IQL) [7], which avoids querying out-of-distribution actions via expectile value regression and advantage-weighted policy extraction. **Model-based RL and planning.** As an exploratory direction we combine a learned ensemble dynamics model in the spirit of MBPO [4] with Monte-Carlo tree search planning [12]. Finally, the POMDP formalism [6] underlies our treatment of SOC-dependent observation masking.

3 Method

3.1 A triage-focused sepsis POMDP

We model triage as a partially observable Markov decision process. The **state** consists of four discretized vital signs (heart rate, systolic blood pressure, oxygen saturation, glucose), three binary treatment flags (antibiotics, vasopressors, ventilation), and the current site of care (asynchronous, ambulatory, facility, ICU); a hidden comorbidity (diabetes) modulates dynamics and is never observed. The **action** is a joint choice of next SOC and the three binary treatments, giving $32 = 4_{\text{SOC}} \times 2^3_{\text{treatment}}$ actions. A **feasibility constraint** $\mathcal{F}(s)$ encodes which treatments are clinically available at each site

Table 1: The base reward function (reward0). Variants modify the boxed magnitudes (terminal scale, treatment, and resource/escalation costs).

Component	Reward / Cost
Death (≥ 3 abnormal vitals)	$-10{,}000$
Discharge (normal vitals, no treatment)	$+10{,}000$
Vital improvement / decline	± 100 per abnormal vital
Unnecessary SOC escalation	$-200 \times \Delta$SOC
High severity without escalation	$-100\times$ SOC gap
Any SOC change	-50
Antibiotics / ventilation / vasopressors	$-10/-60/-40$

(e.g., intensive interventions are unavailable below the ICU), and infeasible action components are masked/clamped at execution.

Partial observability. The observation is a SOC-dependent masked subset of the true state: lower sites of care reveal fewer vitals, mirroring the reduced monitoring available outside intensive settings. This makes site-of-care decisions information-gathering as well as resource-allocation decisions. Figure 1 illustrates one transition.

3.2 Reward variants

Our base reward (**reward0**, adapted from prior work) combines large terminal outcomes with dense shaping (Table 1). Because the milestone diagnostics suggested the terminal scale dwarfed treatment costs — making aggressive treatment near-costless — we designed four variants to probe whether reward design is the bottleneck: **reward1** reduces the terminal scale from $\pm 10{,}000$ to $\pm 1{,}000$; **reward2** adds stronger per-step treatment penalties on top of reward1; **reward3** adds a severity-aware penalty for low-acuity ICU use; and **reward4** adds explicit per-site resource costs.

3.3 Algorithms

We benchmark four families, all sharing the same feasibility-aware action interface.

Rule-based baseline. A deterministic *Heuristic* maps the number of abnormal vitals to a triage-only site-of-care decision, with no learning; it serves as a control whose behavior is, by construction, invariant to the reward.

Online RL. *DQN* [9] and *Double DQN* [13] are value-based baselines; *PPO* [11] is an on-policy clipped policy gradient; *FactPPO* augments PPO with a factorized policy head that decomposes the joint SOC$\times$treatment action into conditionally independent sub-decisions; and *SAC* [3, 2] is a discrete max-entropy actor-critic. We add two SAC variants: *SAC-KL-F* adds the feasibility penalty below, and *SAC-KL-PPO* adds a KL anchor to a frozen PPO reference policy.

Offline RL. *IQL* [7] learns a state value via expectile regression and extracts a policy by advantage-weighted regression, trained on a mixed dataset assembled from the online policies' evaluation trajectories. Our initial offline dataset contained 970k transitions (the online trajectories plus a random-policy supplement for state-space coverage). When we re-ran IQL across the reward variants we had to *recompute* the per-transition rewards under each variant; because the logged observations are SOC-masked while rewards depend on the true state, rewards could not be re-derived from the stored dataset and we instead rebuilt it directly from the trajectory logs (which retain the true abnormal-vital counts), yielding 920k transitions (trajectory-sourced, with the first transition of each episode dropped). Reward-variant IQL results therefore use the 920k rebuild; the two datasets give consistent reward0 behavior. *IQL-KL-F* adds the feasibility penalty.

Feasibility regularizer ("KL-F"). Shared by SAC-KL-F and IQL-KL-F, this penalizes probability mass placed on infeasible actions:

$$\mathcal{L}_{\text{feas}} = \beta\, \mathbb{E}_{s\sim\mathcal{D}}\Big[\textstyle\sum_{a\notin\mathcal{F}(s)} \pi(a \mid s)\Big]. \tag{1}$$

Model-based planning (exploratory). We additionally explored *MBPO + MCTS*: a learned ensemble dynamics model used for Monte-Carlo tree search at decision time, in the spirit of [4, 12]. In a

Table 2: Algorithm comparison at each algorithm's best reward variant (RV), by final mortality, discharge, and infeasible-action rate (mean ± std, lower mortality / higher discharge / lower infeasibility are better).

Algorithm	Best RV	Mortality %	Discharge %	Infeas. %
IQL-KL-F	reward0	$\mathbf{13.6 \pm 6.4}$	$\mathbf{42.4 \pm 7.7}$	0.0 ± 0.0
IQL	reward0	16.0 ± 8.3	41.2 ± 7.2	0.0 ± 0.0
SAC-KL-F	reward0	19.2 ± 9.5	41.6 ± 3.8	0.0 ± 0.0
SAC	reward0	22.0 ± 17.9	39.6 ± 5.2	43.5 ± 28.1
FactPPO	reward0	41.2 ± 14.0	37.6 ± 3.8	0.0 ± 0.0
PPO	reward0	43.2 ± 10.4	40.8 ± 5.9	32.5 ± 41.2
SAC-KL-PPO	reward0	43.6 ± 9.8	40.4 ± 6.2	32.3 ± 40.9
DQN	reward3	47.6 ± 10.5	41.2 ± 5.0	38.2 ± 8.1
Double DQN	reward0	52.8 ± 27.7	32.0 ± 14.7	42.6 ± 24.2
Heuristic	all	93.2 ± 4.1	6.8 ± 4.1	0.0 ± 0.0

controlled study (real environment = training simulator) this loop converged but, with policy and value networks learned from scratch under a limited interaction budget, did not surpass offline IQL; we therefore report it as a method and defer a full evaluation to future work (Section 7.1).

4 Experimental Setup

All algorithms share a common harness with identical environment configuration, fixed deterministic evaluation pools, and a common set of metrics. Online algorithms were trained for 5 random seeds; offline IQL variants were trained on the shared mixed dataset and compared at a converged checkpoint. Every algorithm was run against all five reward variants. We report final-checkpoint **mortality** and **discharge** rates (clinical outcomes), the **infeasible-action rate** (feasibility/safety), and behavioral diagnostics including site-of-care usage and *low-abnormal ICU use* — the fraction of zero-abnormal-vital timesteps spent in the ICU, a proxy for over-triage. All numbers are mean ± standard deviation across seeds. For each algorithm we identify its *best reward variant* (lowest mortality) for the headline comparison, and we additionally report the full algorithm×reward grid.

5 Results

5.1 Quantitative evaluation

Table 2 ranks each algorithm at its best reward variant. Offline **IQL-KL-F** achieves the lowest mortality (13.6%) and highest discharge (42.4%) with zero infeasible actions, followed by IQL and SAC-KL-F. The rule-based Heuristic forms a 93%-mortality floor. Two patterns are immediate: (i) the **feasibility regularizer** reliably eliminates infeasible actions — every "KL-F" method, plain IQL, and FactPPO report 0.0% infeasibility, whereas unconstrained SAC, SAC-KL-PPO. PPO, and the value-based methods place 30–44% of mass on infeasible actions with very large variance; and (ii) **offline methods dominate**: IQL/IQL-KL-F outperform all online learners on mortality.

Figure 4 shows the full 10×5 algorithm×reward grid for mortality and low-abnormal ICU use. The central observation is that the **original reward (reward0) is best for every algorithm except DQN**: reward shaping reliably changes behavior but does *not* jointly reduce mortality and over-triage. Even the strongest method, offline IQL, is not improved by any reward variant — every variant raises its mortality relative to reward0. The Heuristic is, as expected, invariant across the reward axis, anchoring the interpretation.

Training dynamics. Figure 2 shows online learning curves under reward0. As training proceeds, average return rises and mortality falls for the actor-critic methods — confirming the learners are genuinely working rather than collapsing: the SAC family drives mortality down to ≈20% within the first ∼100k environment steps and holds it there, while the value-based methods (DQN, Double DQN) stay high and noisy and FactPPO/PPO sit in between. The fixed Heuristic and the random/no-op references appear as flat baselines. Offline (Figure 3), IQL and IQL-KL-F converge within ∼25k

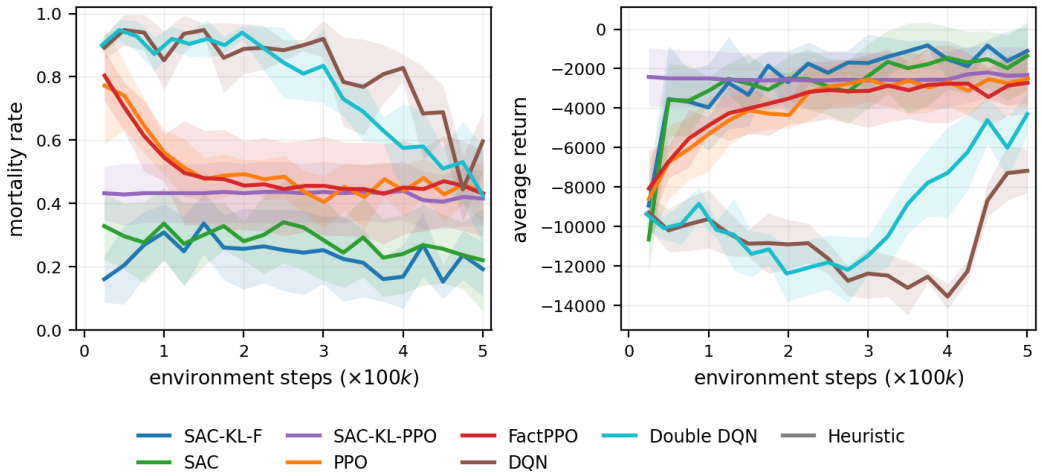

Figure 2: Online training dynamics under reward0 (mean ± std, 5 seeds): mortality (left) and average return (right) vs. environment steps for the seven online methods. Average return rises while mortality falls for the actor-critic methods, confirming the learners are working; the value-based methods (DQN, Double DQN) remain high-mortality and noisy.

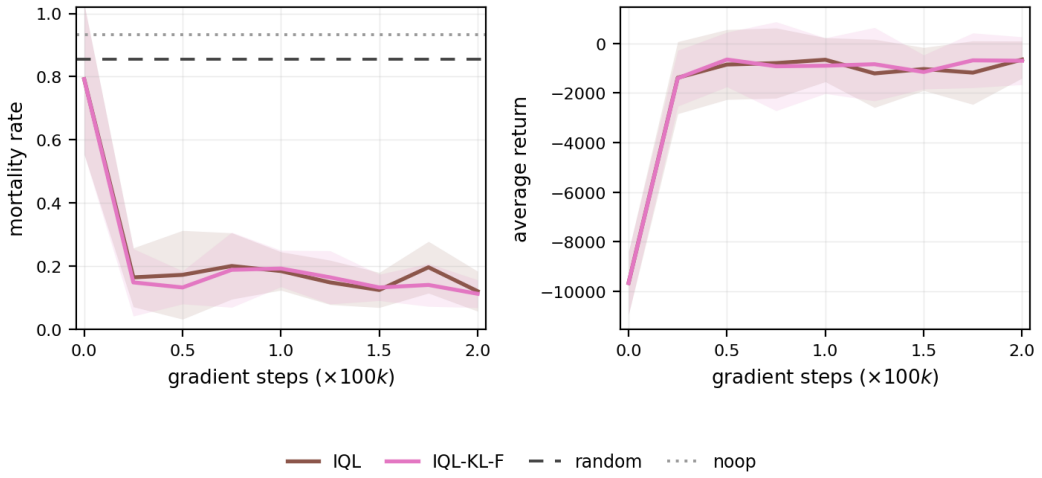

Figure 3: Offline training dynamics under reward0 (mean ± std, 5 seeds): IQL / IQL-KL-F mortality (left) and average return (right) vs. gradient steps. Both converge within ∼25k steps to the lowest mortality of any method, far below the random and no-op references; average return rises sharply from the no-op floor and then plateaus.

gradient steps and remain stable thereafter, reaching the lowest mortality of any method — evidence that the offline dataset's behavioral coverage, not additional online interaction, drives the result.

5.2 Qualitative analysis

Three behavioral trends explain the rankings. First, **actor-critic methods outperform value-based ones**: SAC and PPO achieve substantially lower mortality than DQN and Double DQN. Notably Double DQN is *worse* than DQN in both mortality and variance, suggesting that target decoupling is sensitive to this environment's sparse, terminal-dominated reward structure. Second, **anchoring to a flawed reference hurts**: SAC-KL-PPO inherits PPO's biased behavior envelope rather than improving on it. Third, the **coverage advantage of offline data** is decisive: a ≈0.9M-transition mixed dataset exposes IQL to a broad range of behavior that a 500k-step online learner does not

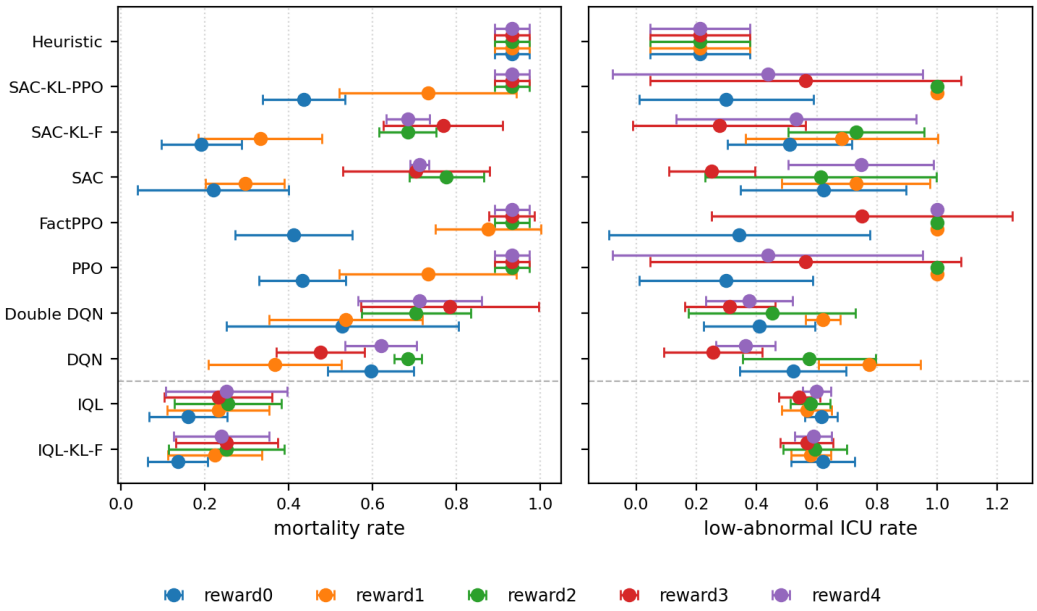

Figure 4: Final-checkpoint mortality (left) and low-abnormal ICU use (right) across all ten algorithms $\times$ five reward variants (mean $\pm$ std, 5 seeds). No variant jointly minimizes both metrics; offline IQL attains the lowest mortality but is likewise not improved by any reward variant.

encounter, and IQL's conservative extraction converts that coverage into the best clinical outcomes while inheriting feasibility directly from the (already-clamped) data.

What the learned policy does. Figure 5 breaks down the best policy's (IQL) treatment choices by site of care and patient acuity, before and after training. The untrained policy (left column) is diffuse, but the trained policy (right column) is highly structured: at zero abnormal vitals it largely withholds treatment at the higher sites of care ("none" at FACILITY/ICU), and as the number of abnormal vitals rises it escalates to combined antibiotic+ventilation and antibiotic regimens, concentrating mass on a small set of aggressive treatments. This is the behavioral signature behind the aggregate numbers: the policy learns an internally coherent escalation rule, yet this coherent policy still leaves substantial high ICU usage at low acuity that no reward variant removes. The action distribution thus makes concrete why the limiting factor appears to lie in the environment-task match: the policy is doing something sensible given the dynamics it was trained on, but those dynamics may not fully support clinically realistic cross-site-of-care triage.

6 Discussion

The results converge on a single, somewhat uncomfortable conclusion. We varied the algorithm across ten methods spanning four families, and we varied the reward across five hand-designed variants; in neither dimension did we find a configuration that produces clinically reasonable triage. The original reward is best for nearly every algorithm, which is strong evidence that the performance ceiling is *not* a reward-specification problem. And the dominance of offline RL trained on logged behavior — rather than any online learner exploring the simulator — suggests the issue is not exploration or optimization either.

What remains is the **environment model** and its match to the task. The original sepsis dynamics are grounded in MIMIC-III, an ICU-derived dataset, while our study extends the simulator to a broader cross-site-of-care triage setting. In that setting, high ICU usage may partly reflect the ICU-centered support of the underlying dynamics rather than purely pathological over-triage. In addition, treatment effects and SOC-specific assumptions remain partly hand-specified, so reward tweaks or algorithm changes may be unable to recover acuity-appropriate behavior if the simulated dynamics make aggressive treatment and escalation broadly favorable. In this light, the feasibility regularizer

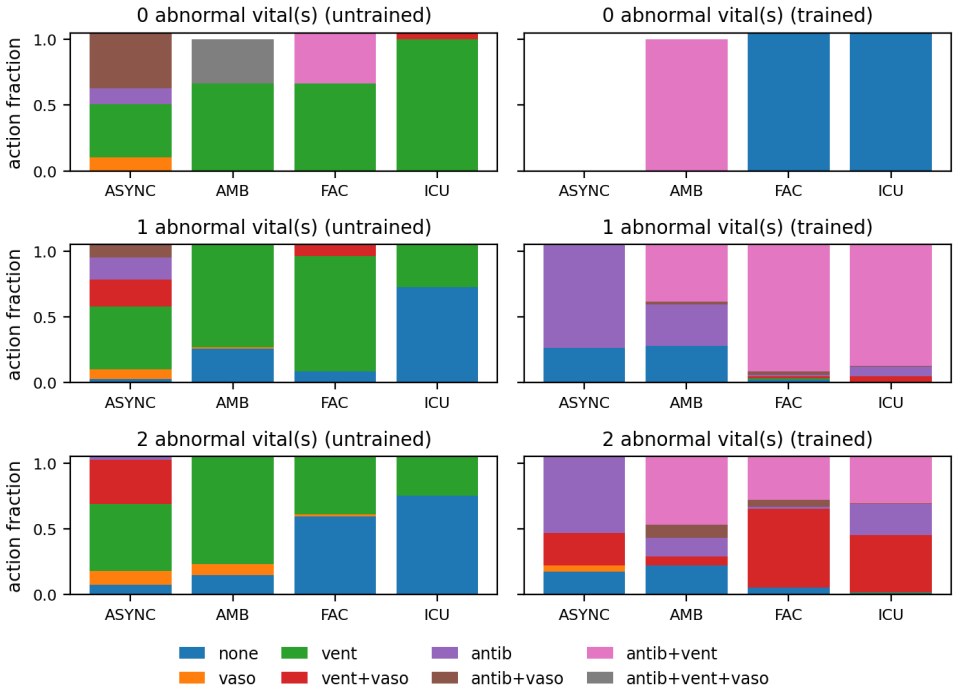

Figure 5: Treatment-action distribution of the IQL policy by site of care (groups within each panel) and number of abnormal vitals (rows), at initialization (left) and after training (right). The trained policy escalates treatment intensity with acuity and site of care, concentrating probability mass on antibiotic / ventilation combinations.

is a clean, deployable win, but the deeper clinical-realism problem lies in validating whether the environment is appropriate for cross-site-of-care triage.

7 Conclusion

We presented a controlled, multi-algorithm, multi-reward study of reinforcement learning for sepsis site-of-care triage in an extended Gumbel-Max SCM POMDP. Offline IQL-KL-F is the strongest method (13.6% mortality, 42.4% discharge, 0% infeasible actions), offline methods beat online ones, and a lightweight feasibility penalty eliminates infeasible actions. Our headline finding is a negative result: because neither reward design nor algorithm choice recovers clinically reasonable cross-site-of-care behavior, the limiting factor appears to be a mismatch between the ICU-derived, partly hand-specified simulator dynamics and the broader triage task. This directly motivates re-grounding and validating the environment model against real clinical data, which is the focus of our future work.

7.1 Limitations and future work

The primary limitation is the simulator-task mismatch. The original dynamics are grounded in MIMIC-III [5], an ICU-derived dataset, while our task extends the environment to cross-site-of-care triage. As a result, high ICU usage by learned policies may partly reflect the ICU-centered support of the underlying dynamics rather than purely pathological over-triage. In addition, some treatment effects and SOC-specific assumptions remain hand-specified, limiting external validity and making it difficult to fully separate reward-design artifacts from dynamics artifacts. Our planned next steps are to (i) re-ground the transition dynamics using empirical patterns from MIMIC-III — currently gated only by data-access credentialing — and sanity-check the simulator by comparing baseline-policy action frequencies against real ICU data; (ii) use the resulting environment to disentangle reward-design from dynamics issues; and (iii) extend the exploratory MBPO+MCTS planner across the reward variants and additional algorithms as stress tests of both simulator and reward.

8 Team Contributions

- **Liane Ozoemelam:** Designed and implemented the triage POMDP simulator (Gumbel-Max SCM extension, site-of-care as state and action, action-space simplification and feasibility constraints); implemented the rule-based Heuristic, Double DQN, and FactPPO; built the trajectory-level evaluation diagnostics; led the poster.
- **Saimai Lau:** Built the shared training/evaluation harness and compute orchestration; implemented and tuned the online learners (DQN, PPO, SAC and SAC variants); assembled the offline dataset and implemented IQL/IQL-KL-F; implemented the exploratory MBPO+MCTS planner; ran the offline algorithm$\times$reward sweep.
- **Yun Dong:** Designed the reward variants (reward0–reward4) and the feasibility and ICU-overuse penalties; ran the reward-variant sweep and built the reward-versioned evaluation and aggregation pipeline; produced the cross-reward diagnostics and analysis.

Changes from Proposal. Our proposal hypothesized that *algorithm choice* would be the primary determinant of success, and the division of labor was organized accordingly. As milestone results showed that algorithm changes alone did not recover good triage, we expanded the scope along two axes that were originally smaller: a systematic *reward-variant* study (Yun) and a substantially *broader algorithm sweep* including offline and model-based methods (Saimai, Liane). The originally planned MIMIC-III grounding was deferred to future work because of data-access credentialing delays. These adjustments shifted effort toward environment- and reward-diagnosis and were necessary to support the project's central finding: the limiting factor appears to be the mismatch between the MIMIC-III-derived, partly hand-specified simulator dynamics and the broader cross-site-of-care triage task, rather than algorithm or reward choice alone.

AI Use Note: Generative AI, including Claude and Claude Code, was utilized to generate scaffolded code, format tables and graphs, pull data from Modal, and concatenate existing results with results from newer runs. RL algorithm logic was implemented largely independently.

References

[1] Kevin Chen, Liane Ozoemelam, and Tanvi Thoria. From acuity to allocation: Learning site-of-care decisions with rl. Stanford CS234 course project, 2025.

[2] Petros Christodoulou. Soft actor-critic for discrete action settings. *arXiv preprint arXiv:1910.07207*, 2019.

[3] 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.

[4] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[5] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[6] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[7] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

[8] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[9] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[10] 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.

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[12] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[13] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

A Reward Variant Details

reward1 reduces the terminal magnitudes of reward0 from $\pm 10{,}000$ to $\pm 1{,}000$. reward2 additionally increases per-step treatment penalties (antibiotics/ventilation/ vasopressors). reward3 adds a severity-aware penalty for occupying the ICU at low acuity. reward4 adds explicit per-site resource costs increasing with site intensity. All variants retain reward0's vital-trajectory shaping and SOC-change costs.

B Implementation Details

Online algorithms were trained for five seeds at 500k environment steps each. The offline dataset (920k transitions for the reward-variant runs; 970k for the original-reward run) was assembled from the online policies' evaluation trajectories with feasibility-clamped actions; IQL/IQL-KL-F were trained for up to 5×10^5 gradient steps and compared at a converged checkpoint. All networks use hidden dimension 256 (128 for PPO/FactPPO), 2–4 layers, and learning rate 3×10. IQL expectile coefficient = 0.8, advantage temperature $_{AWR} = 3.0$, and feasibility regularizer weight = 0.5 for both SAC-KL-F and IQL-KL-F.

Evaluation used fixed per-seed pools of patients, with clinical metrics (mortality, discharge, timeout) and behavioral diagnostics (site-of-care occupancy, action distributions conditioned on site and acuity). The MBPO+MCTS planner used a recurrent ensemble transition model and PUCT-guided tree search with a learned policy prior and value bootstrap.

C Per-Variant Online Training Dynamics

Figures 6–9 show online mortality and average return vs. environment steps under reward variants 1–4 (reward0 appears in the main body, Figure 2; mean $\pm$ std over 5 seeds). Reward shaping changes the learning dynamics — most visibly, the stronger treatment penalties of reward2 and the severity/SOC penalties of reward3–4 push several policies (notably FactPPO, PPO, and SAC-KL-PPO) to collapse toward the high-mortality floor — but no variant improves on the reward0 baseline.

D Per-Variant Action Distributions (IQL)

Figures 10–13 show the IQL policy's treatment-action distribution by site of care and number of abnormal vitals under reward variants 1–4 (reward0 in Figure 5). Across variants the trained policy retains a coherent acuity-driven escalation structure; reward shaping shifts the specific treatment mix but does not remove the over-triage pattern.

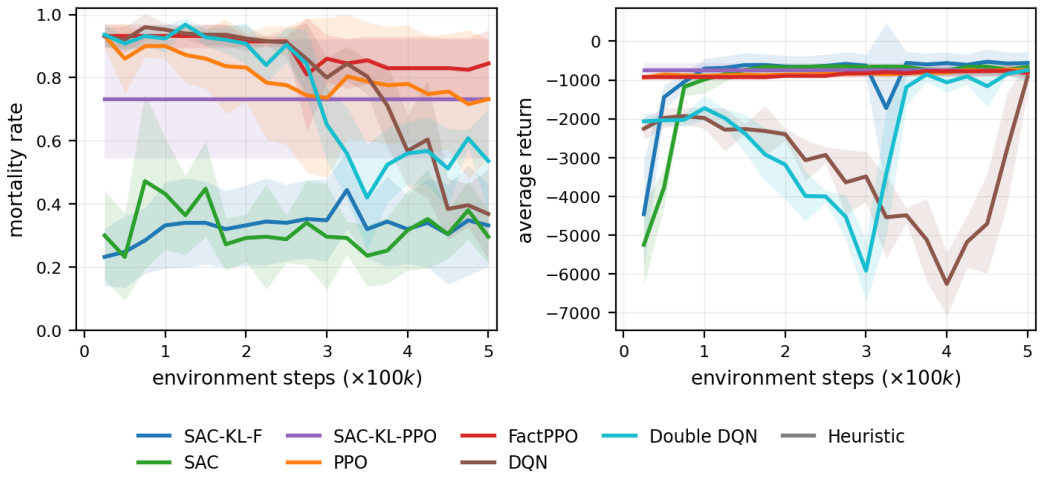

Figure 6: Online training dynamics under reward1 (reduced terminal scale).

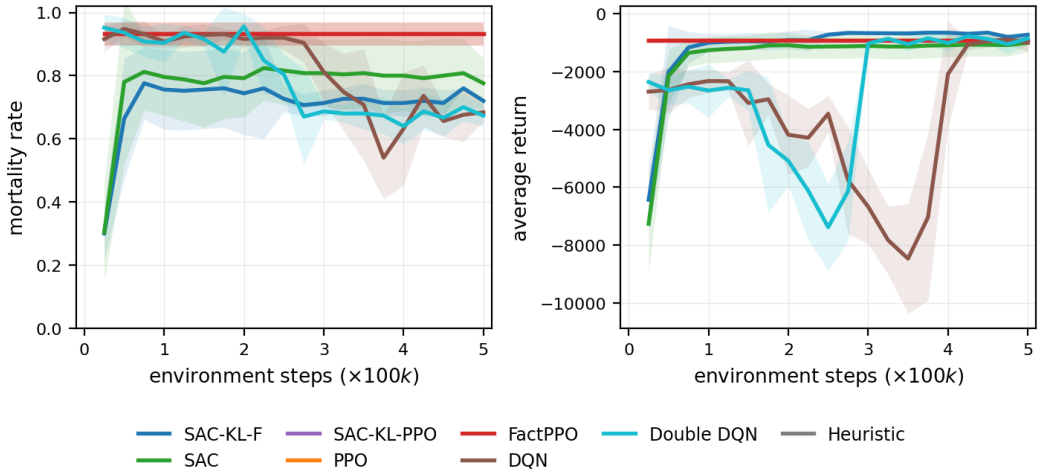

Figure 7: Online training dynamics under reward2 (stronger treatment penalties).

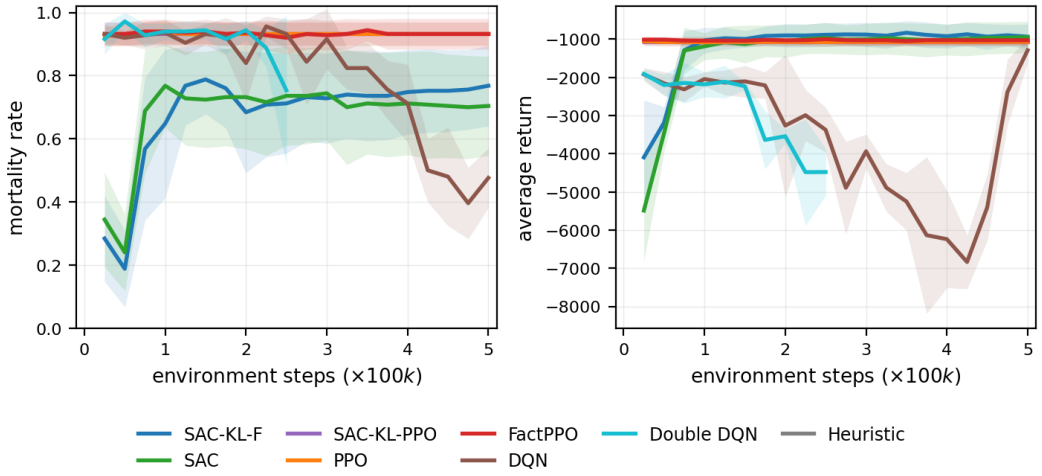

Figure 8: Online training dynamics under reward3 (severity-aware ICU penalty).

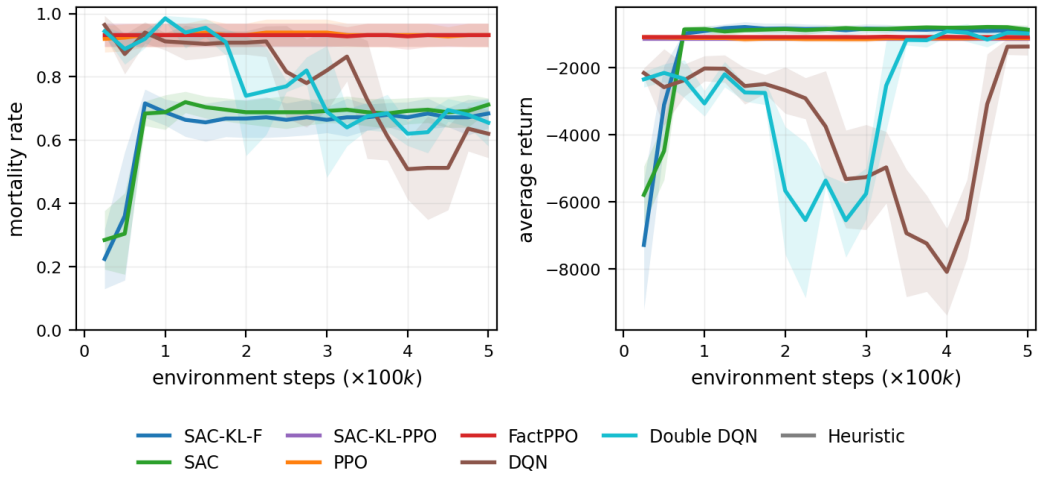

Figure 9: Online training dynamics under reward4 (SOC resource costs).

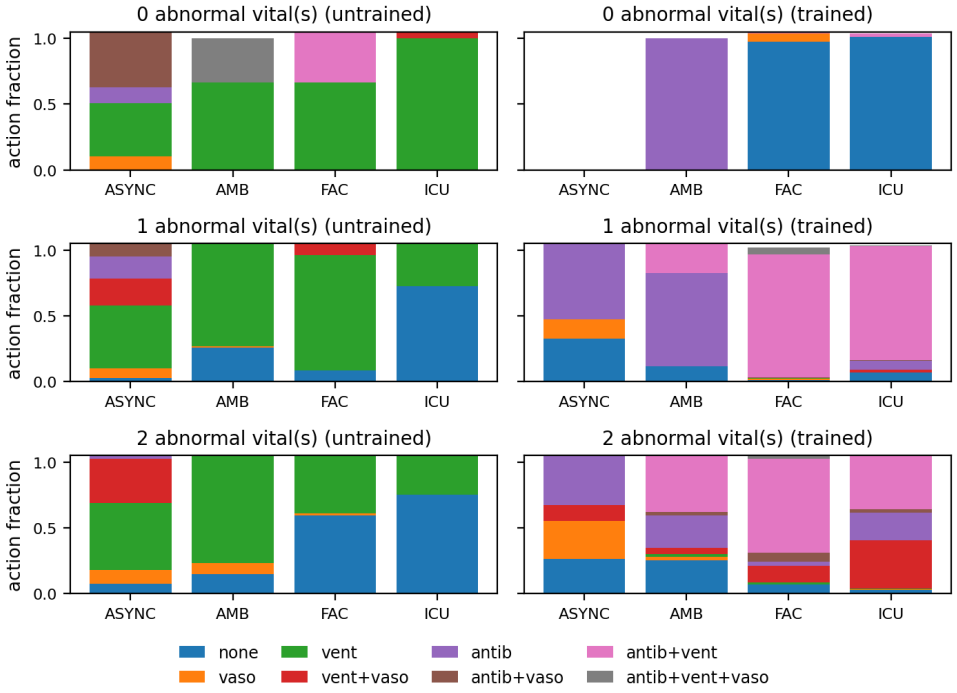

Figure 10: IQL action distribution under reward1.

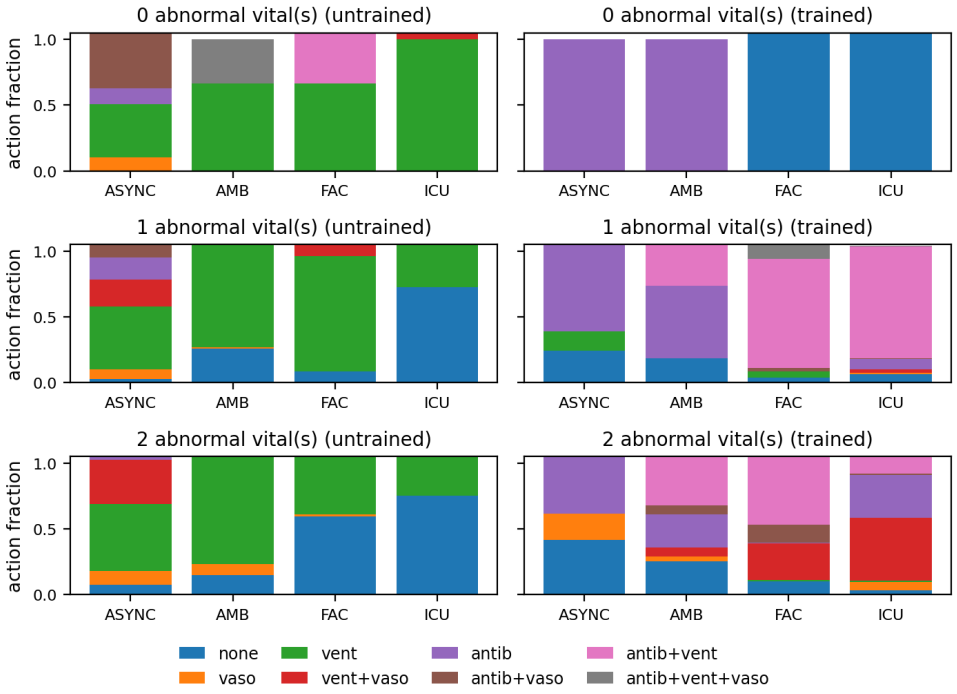

Figure 11: IQL action distribution under reward2.

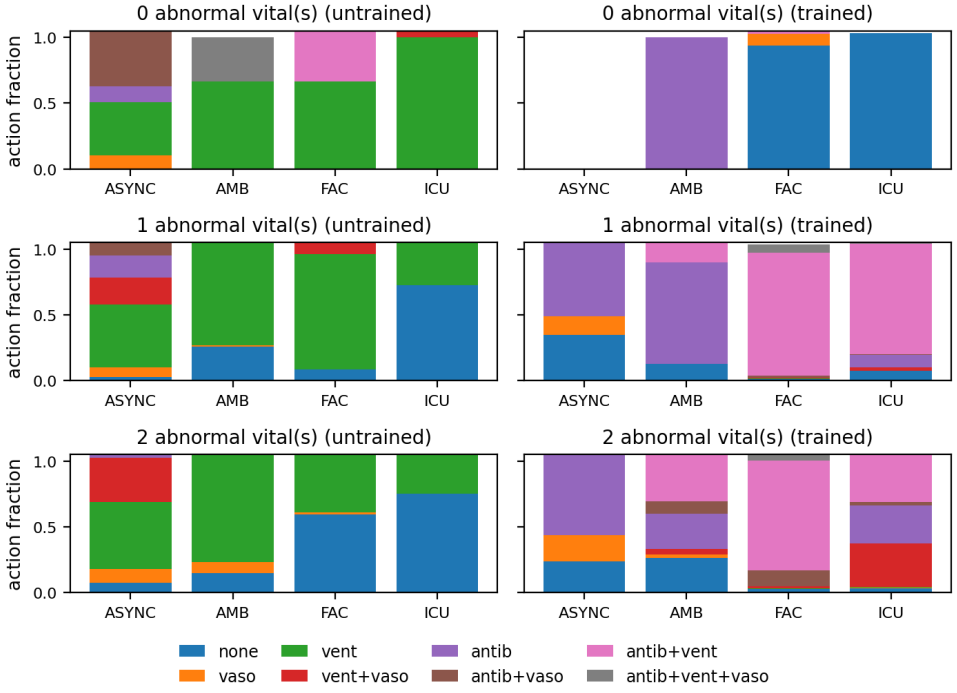

Figure 12: IQL action distribution under reward3.

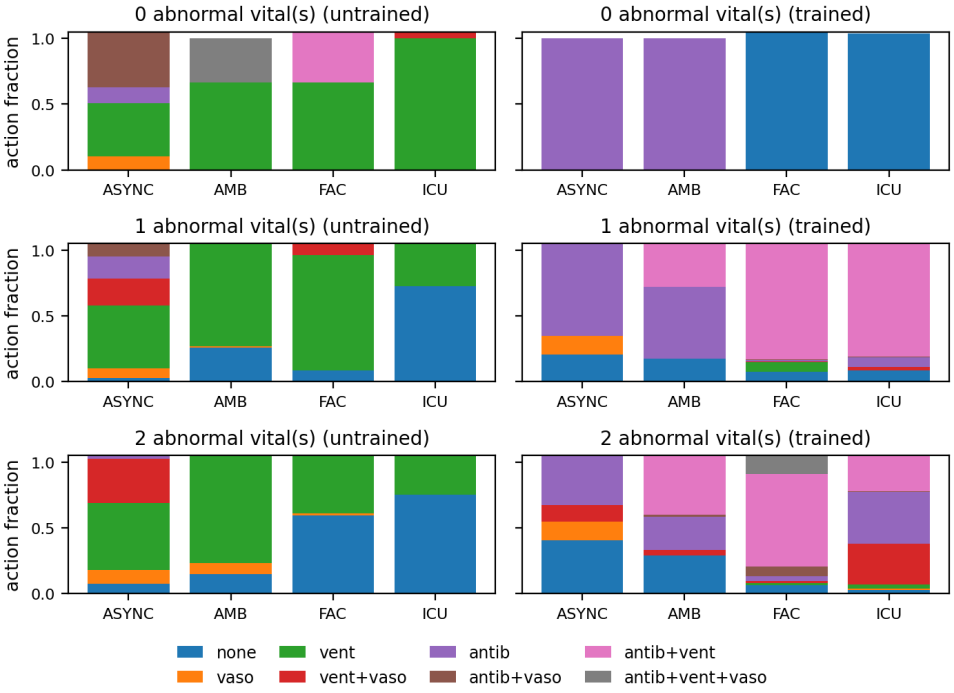

Figure 13: IQL action distribution under reward4.