EHRGym: A Diagnostic Benchmark for Offline Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

A low offline-RL score on clinical-style logs can reflect little room to improve, missing information, or failed learning. EHRGym separates these possibilities in one synthetic ventilator task with hand-specified responses and rewards. Its reference ladder runs from the exactly computed best constant setting to a hidden-state oracle. Planners that know the simulator probe what past settings, timing, and measurements can support. A clock-only plan follows a fixed schedule of settings indexed by time and uses no measurements. On varied logs, this plan reaches about a quarter of the way from floor to ceiling. DQN/FQI and MedDreamer see the full record yet stay near the floor. With hidden-state access, DQN/FQI reaches more than halfway, exposing difficulty learning from the partial record. Under repetitive logging, CQL and Decision Transformer often copy starting settings and score below their stochastic behavior policy. Researchers can use information-matched references to interpret low scores and inspect chosen actions for logging habits.

1 Introduction

Why is offline RL for treatment decisions hard to evaluate? A developer trains an offline reinforcement learning (RL) method on ventilator logs and gets a low score (Komorowski et al., 2018; Raghu et al., 2017). Did changing settings offer little value? Did the record omit information needed to act? Or did the method fail to learn from that record? The record includes measurements, their times, and past settings. Logged care cannot reveal outcomes under unchosen settings, so these explanations remain tangled (Gottesman et al., 2019; 2018).

Why a new benchmark? A diagnosis needs three ingredients. Exact floor and ceiling values establish whether improvement was possible. Reference planners restricted to what a policy can see test what its record supports. Controlled logging and clinically informed measurements show how data generation shapes learning. No compared benchmark provides all three (Table 1 and Section 2).

EHRGym. EHRGym uses one synthetic ventilator task with measurements informed by MIMIC-IV aggregates (Johnson et al., 2023; 2024). Its treatment responses and rewards are hand-specified. A reference ladder is a set of scores from a floor to a ceiling (Figure 1). The best constant setting defines the exactly computed floor. An oracle that sees the hidden state defines the exactly computed ceiling. Together, they show whether improvement was possible in the simulator. A clock-only plan changes settings at preset times without reading measurements. Its score shows what time alone achieves. Known-model planners see past settings and timing, with or without measurements. A learner below a reference using the same or less information leaves learnable value unused; simulator-aware references show what that information can achieve, not its optimum. Researchers can vary measurement acquisition, decision timing, and logging while keeping treatment responses and rewards fixed (Section 3.2).

What we find. With varied logging, the clock-only plan reaches about a quarter of the floor-to-ceiling span. DQN/FQI and MedDreamer see the full patient record but stay near the floor. With hidden-state access, DQN/FQI reaches more than halfway (Figure 1 and Section 5.1). With repetitive logging, CQL and Decision Transformer often repeat the setting each episode starts with. Under deterministic deployment, they score below the stochastic policy that produced their data (Section 5.2).

Why it matters. Methods researchers should report information-matched references alongside scores on clinical-style logs. They should compare chosen actions with the behavior policy's habits. In this

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

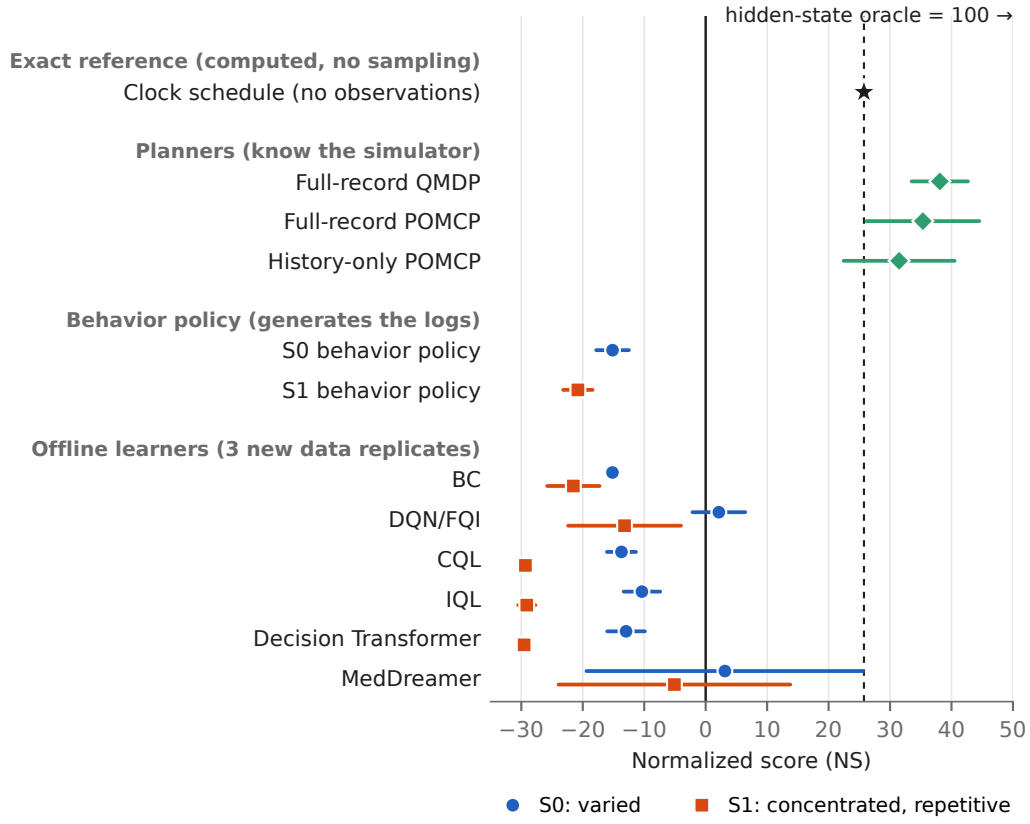

Figure 1: **Respiratory normalized scores under severity-driven measurement (M1) and held-action timing (T1).** S0 means varied behavior-policy choices; S1 means concentrated, repetitive choices. The dashed line marks the best-found clock schedule, evaluated exactly and a lower bound on the best clock-only schedule. The hidden-state oracle lies off-axis at 100 NS. Learner intervals use three data replicates; planner and behavior-policy intervals use evaluation episodes.

synthetic task, learning from partially observed, observation-driven logs remains the open problem despite demonstrable room for improvement.

EHRGym contributes a reference ladder with exact floor and ceiling values, planners restricted to different records, and controlled measurement, timing, and logging conditions. C1 (attainable-value gap): record-based learners trail references using the same or less information. C2 (behavior-policy robustness): under repetitive logs, deterministic learners copy the behavior policy's most likely action and lose value. C3 (information use): planners show the value achieved with clock time, action and timing history, and measurements.

2 Related Work

Existing benchmarks cover parts of this setting (Table 1). Exact values show possible improvement; record-limited planners show achieved value, and both help interpret low learner scores. EHR-informed acquisition supplies recognizable measurements; controlled logging tests repetitive data. EpiCare tests offline learners with varied data sources and a nonprivileged symptom-based policy, but reports neither a history-restricted planner nor an exact finite-horizon oracle (Hargrave et al., 2024). DTR-Bench varies noise and missingness and evaluates learners including recurrent variants, but does not report a history-restricted planner (Luo et al., 2024). ICU-Sepsis provides an exact fully observed tabular MDP reference, not a history-restricted planner (Choudhary et al., 2024). MedGym includes irregular timing, clinical-data-derived scenarios, offline learners, and behavior-policy-generated data,

Table 1: **Reported benchmark ingredients.** Clinical denotes a healthcare decision task; Offline RL denotes log-trained baselines; Timing changes decision opportunities; Actions reports selected actions. EHR acquisition uses clinical-data-informed observations; Logging varies the behavior policy; Exact denotes model-computed values; History denotes a planner restricted to policy-available information. EHR acquisition, logging, and exact values jointly support the Respiratory diagnosis. Marks summarize the cited reports, not all possible implementations.

	Shared foundations				Diagnostic ingredients			
Benchmark	Clinical	Offline RL	Timing	Actions	EHR acquisition	Logging	Exact	History planner
EHRGym	✓	✓	✓	✓	✓	✓	✓	✓
EpiCare	✓	✓	×	✓	×	✓	×	—
DTR-Bench	✓	×	—	—	×	—	—	—
ICU-Sepsis	✓	×	×	—	—	—	✓	—
MedGym	✓	✓	✓	✓	∼	∼	—	—
B4MRL	×	✓	—	—	×	✓	—	—
D4RL	×	✓	—	—	×	✓	×	—
OGBench	×	✓	—	—	×	✓	—	—

✓ = yes; ∼ = partial; × = no; — = not reported, not absent. Marks reflect each paper's own description.

but reports no matched logging variation, history-restricted planner, or exact value (Wang et al., 2026). B4MRL probes biased offline data and imperfect simulators; D4RL provides normalized scores; OGBench tests capabilities outside healthcare (Linial et al., 2024; Fu et al., 2020; Park et al., 2025). EHRGym combines these diagnostic ingredients in synthetic Respiratory (Appendix O).

Off-policy evaluation from logs can be biased when the behavior policy uses observations unavailable to a learner (Tennenholtz et al., 2020); hidden confounding also motivates dedicated offline RL methods (Pace et al., 2023). EHRGym supplies a controlled setting in which the behavior policy uses only available observations and the previous action. A time-step-only learner cannot see the measurements used to choose logged actions, so confounding is consistent with its shortfall, not identified as its cause. Learners that see the full available record receive the behavior policy's inputs yet remain below the clock schedule, pointing to difficulty learning control from partially observed logs (C1; Section 5.1; Appendix F.1).

Observation-history imitation can copy previous actions (Wen et al., 2020); BCQ, TD3+BC, CQL, and IQL constrain offline learning toward recorded behavior in different ways (Fujimoto et al., 2019; Fujimoto & Gu, 2021; Kumar et al., 2020; Kostrikov et al., 2022). Repetitive logs make copying a plausible failure, but copying alone does not show whether a learner loses value relative to the policy that generated those logs. In EHRGym's controlled M1/T1/S1 condition, reset, persistence, and deployment tests locate that loss: CQL and Decision Transformer repeat the starting action and score below the stochastic behavior policy (Table 23).

3 Benchmark Design

Figure 2 shows the benchmark workflow.

3.1 EHR-informed observations and synthetic decisions

Each EHRGym profile specifies a partially observed decision process, an action inventory, and a reward law. For the present Respiratory task, aggregate MIMIC-IV version 3.1 analyses supply measurement-rate inputs and a clinical measurement vocabulary. A train-only rank-four probabilistic principal-component fit supplies correlated observation background (Tipping & Bishop, 1999). The input applies oxygenation, blood-pressure, and urine-output mappings. The available record contains measurements and their timing, with past actions; no patient-level record enters this manuscript package (Appendix B). The response process, treatment effects, terminal events, and rewards are synthetic, rather than estimated as effects of observed care. Hidden burden, response factors, exposure memory, pending treatment phases, and previous action govern stabilization and critical events. Aggregate rates and a background fitted on training data inform emitted measurements. M0

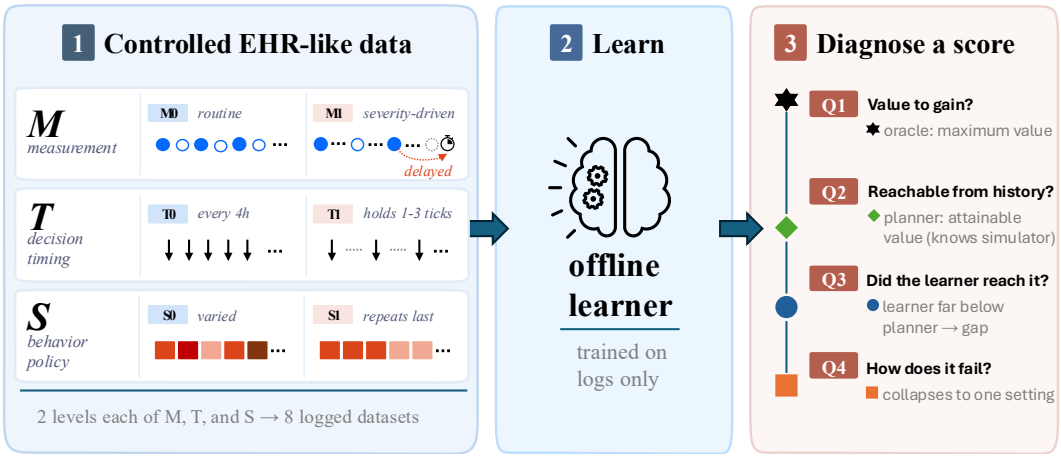

Figure 2: **Controlled data, learning, and score diagnosis.** Panel 1 makes eight logged datasets by varying measurement M (M0 routine; M1 severity-driven with delayed results), decision timing T (T0 every 4 h; T1 holds 1–3 ticks), and behavior policy S (S0 varied; S1 repeats the last setting). Panel 2 trains learners on logs; panel 3 maps Q1 and Q3 to C1, Q2 to C3, and Q4 to C2. Scores are schematic; Figure 1 reports measured values.

follows aggregate acquisition rates. M1 adjusts acquisition using severity, availability, recency, and previous action. Unobserved values and hidden states remain hidden (Appendix B).

Respiratory has 12 supported PEEP/FiO_2 combinations. The lowest normalized-intensity action is the lowest supported combination; it is not the absence of oxygen or ventilation. Physical time advances in four-hour ticks for at most 11 ticks, with $\gamma = 0.99$. Actions are held over a decision interval. The per-tick reward is $+1$ at stabilization, -1 at a critical event, and $-0.05I(a)^2$ treatment burden, where $I(a)$ is normalized action intensity. Held-tick rewards use physical-time discounting. The evaluator can access the synthetic hidden state to calculate exact values, while ordinary policies receive available observations and their own previous actions. Decision Transformer also consumes realized interval reward after each transition to update its return target. Appendix N lists source records.

3.2 Controlled M/T/S conditions

A $2 \times 2 \times 2$ design generates eight offline tasks per profile. M0 uses the base empirical acquisition schedule. M1 shifts acquisition toward profile-relevant measurements and changes whether new results are available at the current decision. Both retain partial observation of the same hidden response process. T0 permits an action every tick. T1 holds an action for one, two, or three ticks according to hidden burden. Holds stop at the task horizon, and rewards use physical-time discounting. The elapsed duration is observed, so T1 changes when actions can change and carries severity information. S0, the varied behavior policy, and S1, the concentrated and repetitive behavior policy, generate the *offline training and selection logs*. The fixed behavior policy is an authored observation-driven heuristic, not a fit to MIMIC action frequencies. S0 and S1 change its temperature and previous-action persistence while retaining positive action support (Appendix I). Its concentration parameter changes from $\tau = 1.0$ to 0.35, and its previous-action persistence from $\kappa = 0.10$ to 0.75. S1 therefore retains positive support while jointly increasing concentration and persistence. The evaluation response, reward, and action laws are identical across S0/S1.

We evaluate all eight M/T/S conditions for Respiratory. M changes acquisition without changing hidden responses; T changes action opportunities and reveals informative elapsed duration; S changes logs under the same evaluation process. A separate experiment varies behavior-policy concentration and persistence.

Table 2: **Eight-condition Respiratory suite.** Every M/T/S combination is included; S changes logged data, not the evaluation mechanism.

Axis	Level 0 versus level 1	Diagnostic question
M	Base acquisition versus severity/availability/recency/action-adjusted acquisition and result availability	How does measurement process affect learning?
T	Action each tick versus burden-dependent one-to-three-tick hold	How do decision opportunities and observed hold durations affect learning?
S	Varied versus concentrated and repetitive behavior policy; both retain positive support	How does logged-action distribution affect learning?

3.3 Using EHRGym

Choose a condition, train and select on separate logs, then evaluate by simulator rollout. Report timing-specific NS, replicate intervals, actions, behavior-policy support, and reward components (Appendix A).

4 Methods

4.1 Exact values and planners

Let $V_p(a^*)$ be the exact value of the best constant legal action and V_p^* the exact finite-horizon hidden-state oracle under profile p's default M1/T1 response, reward, and timing laws. Define

$$H_p = V_p^* - V_p(a^*), \qquad G_p = V_p(\pi_{\text{pub}}) - V_p(a^*), \qquad \rho_p = G_p/H_p \quad (H_p > 0). \tag{1}$$

H_p is a constant-to-hidden-state-oracle opportunity gap. G_p is the value achieved by a full-record planner that knows the simulator. Its estimate comes from evaluating that planner on simulated episodes.

For return V under timing T, NS uses the exact best constant and timing-specific hidden-state oracle:

$$\text{NS}_T(V) = 100\frac{V - V_{\text{const}}}{V_{\text{oracle,T}} - V_{\text{const}}}, \quad V_{\text{const}} \simeq -0.5827, \quad V_{\text{oracle,T1}} \simeq +0.0435. \tag{2}$$

This post hoc transformation preserves within-timing rankings and can fall outside 0–100. The best constant is invariant across M/T/S, while T0 and T1 have different oracle denominators; S changes logs, not these exact reference values (Appendix G).

At M1/T1, the exact references define NS and the best-found clock schedule uses physical tick alone. Known-model planners compare history-only and full-record inputs (Figure 1; Table 4; Appendix D.1).

4.2 Learner panels and selection

An initial screen evaluated 12 EHRGym methods on one logged dataset per M/T/S condition. Appendix C lists them and their objectives. All policies use the available record and legal-action mask. BC samples from a masked stochastic policy. The other eleven entrants execute deterministic masked argmax actions when evaluated; a prespecified S1 control also tests sampled deployment of CQL and Decision Transformer.

We reran six learners on new data, with three logged-data replicates per condition: BC, DQN/FQI, CQL, IQL, Decision Transformer, and MedDreamer. They span imitation, value learning, conservative and implicit learning, sequence modeling, and a world model. All six receive the available record before each decision. Decision Transformer also receives realized reward after each interval (Appendix C). Each method and suite condition has three independent logged-data replicates (1–3), with 16,384 training episodes and 2,048 offline-selection episodes per replicate. Within each condition and replicate, methods share logged data and the same normalization statistics. Configurations are unchanged from the initial screen. Checkpoints are selected by a method-specific held-out objective, never evaluation return. The selected checkpoints are directly evaluated on 8,192 simulator episodes per fit.

4.3 ESTIMATES AND UNCERTAINTY

For an M1/T1 suite condition, the learner estimate averages three independent logged-data replicate means. Two-sided 95% intervals use a t distribution with two degrees of freedom and the sample standard deviation of those three means divided by $\sqrt{3}$. All replicate intervals are nominal and unadjusted unless stated otherwise. Across all eight conditions, each replicate contributes either an equal-cell mean or a difference between axis levels. The three replicate summaries then determine the interval.

QMDP and POMCP use the declared synthetic model and available histories (Silver & Veness, 2010). Only the evaluator's hidden-state oracle uses hidden-state information. Learner intervals use logged-data replicates; planner and behavior-policy intervals use evaluation episodes. The POMCP comparison uses 1,024 shared environment seeds; full-record QMDP and the full-record POMCP action audit use separate runs (Table 4; Appendix B). We inspect the actions learners choose by replaying each selected checkpoint and recording action frequencies and behavior-policy support (Appendix K).

4.4 PLANNERS AND THE BEHAVIOR-POLICY EXPERIMENT

At M1/T1, history-only POMCP knows the synthetic model and sees clock time, past actions, durations, and continuation. Full-record POMCP also sees measurements and their times; neither sees the hidden state (Appendix E.1).

The 2×2 M1/T1 follow-up varies behavior-policy concentration and previous-action persistence separately. BC, CQL, and Decision Transformer use three paired data blocks with separate training and selection roles. The environment, reward, and selection rules remain fixed; Appendix J.3 gives settings, fit counts, and interval methods.

5 RESULTS

5.1 C1: ATTAINABLE-VALUE GAP

Learners given the hidden state learn much of the task value (Table 3). On varied logs, record-only DQN/FQI scores 2.1 NS in the main confirmation and MedDreamer scores 3.1 NS (Table 4); both trail the best-found clock schedule, which uses no observations. The exact belief adds a small paired gain to DQN/FQI but leaves its score below that schedule. Simple fitted-Q learners show the same ordering: hidden state above belief above time step. Yet the time step is not a Markov state; fitted-Q Bellman backups over it alone do not recover the clock schedule. This result cannot by itself isolate what is learnable from the logs. The clock schedule and planners use the simulator model; to test logs alone, we searched schedules from five independent S0 logs. Every action in the best-found schedule appears at every time step in all five logs, yet tick-only fitted-Q reaches 1.6 NS by exact evaluation and logged-means search 13.2 NS (Appendices F.1 and F.2). Importance-weighted (WPDIS) search reaches -20.9 NS, with final effective sample size 1.5 of 16,384 training episodes; record-based DQN/FQI and MedDreamer score 2.1 and 3.1 NS. Thus EHRGym distinguishes known-model value (clock schedule 25.8 NS; planners 31.5–41.3 NS) from these log estimators' mean scores (at most 13.2 NS). Observation-driven behavior-policy choices are consistent with the logged-means shortfall; these results do not identify the best schedule learnable from logs.

Is the gap a training artifact? Matched DQN/FQI controls test checkpoint choice and learning rate. A 2×2 visibility-by-behavior-policy control tracked 18,000 updates (Figure 3; Appendix F.4). Mean NS peaks by 3,000 updates in all four cells, then falls to -15.1 to -25.0 by 18,000: training is unstable with or without hidden-state access. At every saved update, available-record means stay below the 25.8 NS clock schedule, peaking at 2.4 NS on S0 logs and 6.7 on uniform logs; hidden-state means peak at 61.5 and 62.1 NS. Optimistically choosing each fit's best curve checkpoint, unavailable offline, gives at most 12.6 NS with the available record, while every hidden-state fit reaches at least 58.1 NS. Simulator-evaluated tuning retains the default S0 DQN/FQI learning rate (7.4 NS; Appendix L); tabular fitted-Q, without neural optimization, shows the same ordering (Table 3). The main confirmation's 2.1 NS, belief control's 0.01 NS, and visibility control's -0.77 NS use separate record-only log replicates and all lie near the floor. In these controls, C1 follows learner

Table 3: **What the learner sees at M1/T1 on S0 logs.** Entries are mean NS [95% replicate t interval]; — means the input was not tested. Simple fitted-Q uses five log/fit replicates; DQN/FQI record and belief entries use three paired belief-control replicates, while its hidden-state entry uses three separate matched-visibility replicates (Appendices F.1, F.5, and F.4). The DQN/FQI record-only value (0.01) is the paired belief-control baseline; Table 4 gives the main confirmation estimate.

Input	Simple fitted-Q NS	DQN/FQI NS
Time step only	$1.37\ [-3.99, 6.74]$	—
Time step + exact belief	$9.43\ [5.20, 13.67]$	—
Available record	—	$0.01\ [-16.40, 16.43]$
Available record + exact belief	—	$5.05\ [-10.93, 21.03]$
Time step + hidden state (simple)	$41.86\ [39.11, 44.61]$	—
Record + hidden state (DQN/FQI)	—	$57.36\ [51.77, 62.94]$

References (NS): clock schedule 25.8; history-only POMCP 31.5; full-record POMCP 35.3; oracle 100.

visibility rather than checkpoint timing or the tested learning rate. Increasing logged training data from 16,384 to 262,144 episodes gave no detectable gain for DQN/FQI, CQL, or MedDreamer under fixed optimization budgets (Appendix F.3).

DQN/FQI exceeds the S0 behavior-policy point score at M1/T1; MedDreamer has a higher point estimate, but its replicate interval overlaps that score. These unpaired point comparisons show improvement over the policy that generated S0 logs while leaving the planner gap (Table 4; Appendix F).

5.2 C2: Behavior-policy robustness

Under S1, deterministic CQL, IQL, and Decision Transformer score below the stochastic behavior policy (Table 4). The persistence sweep and deployment controls show that deterministic choice reproduces the policy's repeated-action mode, while sampled deployment shows no detectable deficit against that policy in the tested controls (Appendices J.2 and J).

S1 gives a known answer for behavior-policy copying: its unique conditional mode is the previous action, so greedy copying repeats reset action 0 (exact return -0.7669, or -29.4 NS). CQL, IQL, and Decision Transformer score -29.3, -29.1, and -29.5 NS (Table 4), matching this prediction. Avoiding this failure means exceeding the stochastic behavior policy's -20.8 NS under deterministic deployment; CQL with weaker conservatism ($\alpha = 0.1$) reaches -15.9 NS, though simulator-based selection makes this optimistic (Appendix L).

The S1 loss accompanies repeated selection of the lowest supported PEEP/FiO_2 pair, action 0. In 8,192 M1/T1/S1 episodes per fit, CQL, IQL, and Decision Transformer select it in 99.5%, 99.6%, and 99.9% of decisions; full-record POMCP selects it in under 0.1% in both evaluated runs (Figure 4b; Appendix K). Across S1 conditions, these learners fall 0.18–0.19 below the best constant (Figure 4); S1 reward components show treatment-cost savings outweighed by simulated clinical losses.

The M/T/S axis effects are reported in Appendix H.

Persistence or concentration? Across three paired M1/T1 blocks, we varied low/high concentration and low/high persistence for BC, CQL, and Decision Transformer. Higher persistence makes CQL and Decision Transformer select action 0 in at least 99.6% of decisions at both concentrations; return contrasts are about -0.11 with simultaneous intervals below zero (Table 23). BC, concentration, and interaction intervals span zero; clinical losses exceed treatment-cost savings (Appendix J.3). Persistence drives the tested effect for these two methods.

Resetting the starting setting shows that CQL and Decision Transformer repeat that setting rather than action 0 specifically. Sampling deployed actions improves their returns; removing previous-action input and weakening CQL's penalty reduce repetition in the tested controls (Table 23; Appendices J.5, J.4, and L). The persistence sweep and full estimates are in Appendix J.2.

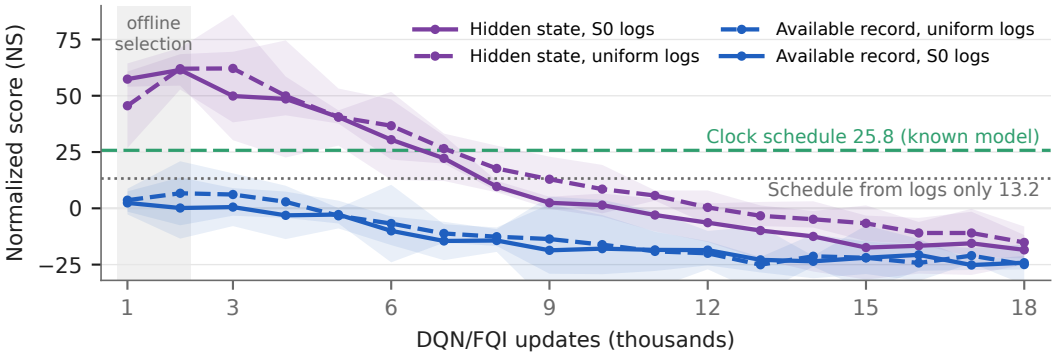

Figure 3: **DQN/FQI checkpoint curves at M1/T1.** Curves use a separate 1,024-episode set per checkpoint that did not select checkpoints, show mean NS over three new log/fit replicates per cell with 95% replicate t bands, and shade the offline-selected 1,000 or 2,000 updates. Horizontal lines mark the exactly evaluated best-found clock schedule and the best log-selected schedule (mean of five log replicates).

Table 4: **Respiratory M1/T1 leaderboard.** NS uses Eq. 2. Learner 95% intervals use three logged-data replicates; planner 95% intervals use episodes; behavior-policy intervals describe direct evaluation episodes; exact values have no interval. The POMCP shared-seed run uses 1,024 common environment seeds; the 41.3 NS full-record POMCP and 38.0 NS full-record QMDP each use a separate run (Section 4.3).

Policy or reference	S0 NS [95% interval]	S1 NS [95% interval]
BC	-15.2 $[-15.5, -14.9]$	-21.5 $[-25.8, -17.3]$
DQN/FQI	2.1 $[-2.1, 6.4]$	-13.2 $[-22.4, -4.0]$
CQL	-13.7 $[-16.1, -11.3]$	-29.3 $[-30.0, -28.7]$
IQL	-10.4 $[-13.3, -7.4]$	-29.1 $[-30.5, -27.7]$
Decision Transformer	-13.0 $[-16.0, -9.9]$	-29.5 $[-29.9, -29.1]$
MedDreamer	3.1 $[-19.4, 25.7]$	-5.1 $[-23.9, 13.8]$
S0 behavior policy	-15.2 $[-17.8, -12.5]$	
S1 behavior policy	-20.8 $[-23.2, -18.4]$	
Exact best constant	0.0 (exact)	
Clock schedule	25.8 (exactly evaluated)	
History-only POMCP (shared-seed run)	31.5 [22.5, 40.5]	
Full-record POMCP (shared-seed run)	35.3 [26.2, 44.5]	
Full-record POMCP (separate run)	41.3 [32.0, 50.6]	
Full-record QMDP (separate run)	38.0 [33.4, 42.6]	
Exact hidden-state oracle	100.0 (exact)	

5.3 C3: Information use

At M1/T1, best constant scores 0 NS, clock schedule 25.8, history-only POMCP 31.5, full-record POMCP 35.3, and hidden-state oracle 100 (Figure 1; Table 4). The paired POMCP measurement contrast is inconclusive at $+0.02$ return $[-0.02, +0.07]$; the paired QMDP contrast is positive at $+0.0500$ $[+0.0259, +0.0741]$ ($+8.0$ NS; Appendix Table 9). Thus the POMCP point difference alone does not establish a measurement benefit. Clock time supplies much of the achieved value: history-only POMCP scores about 6 NS above the clock schedule and full-record POMCP about 4 NS above history-only POMCP. These point differences do not decompose information value.

In the timing-law QMDP control, planning from past settings and timing retains most of the gain even when hold durations no longer depend on hidden burden; the measurement increment at that scale remains inconclusive (Appendix E.3).

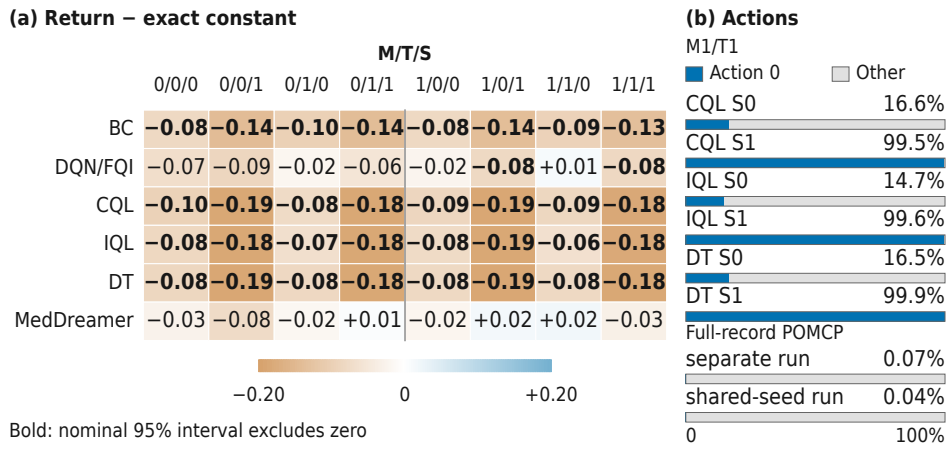

Figure 4: **Respiratory learner results and actions.** (a) Mean raw return minus the exact best constant across eight conditions; column codes list measurement (M), decision timing (T), and behavior policy (S) in that order, so 0/1/0 means base acquisition, burden-dependent held actions, and behavior-policy variation. Bold cells have replicate intervals excluding zero. (b) M1/T1 action frequencies show concentration on the lowest supported setting under S1; the two full-record POMCP bars represent two evaluation runs. Exact cell values and uncertainty are in Appendix F.

6 What EHRGym reveals

When a learner scores low, EHRGym lets developers check three things. First, hidden-state learners find value, while record-based learners trail a clock schedule that needs no observations; logged schedule search exposes estimation difficulty (C1; Section 5.1). Second, planners show achieved value with clock time, action and timing history, and the full record (C3; Section 5.3). Third, under repetitive logs, deterministic CQL and Decision Transformer copy the behavior policy's most likely action, while their sampled deployment shows no detectable deficit against that policy in the tested controls (C2; Section 5.2). Developers should report chosen actions, behavior-policy support, and reward components alongside returns.

7 Limitations

Only Respiratory is evaluated. MIMIC-IV aggregates inform measurements; authored responses and rewards establish no clinical utility or treatment effects. S1 is designed, not fitted to observed actions; adjacent-pair repetition does not establish continuous persistence. Learner and hyperparameter tests are limited; matched controls cover only DQN/FQI, and the belief-input control uses only S0 logs. The time-step-only fitted-Q learner uses a non-Markov input and lacks measurements used to choose logged actions; its shortfall does not identify learnability or confounding. Planners show achieved value, not an optimum or finite-log learnability; convergence and paired POMCP measurement contrast remain unresolved. Control limits: Appendices B, E.3, F.3, J.3, J.5, and L.

8 Conclusion

Why does an offline learner score low? In a synthetic ventilator task with hand-specified responses and rewards, EHRGym shows exact room for improvement: a clock schedule achieves value without measurements and exceeds tested record-based learners, while hidden-state DQN/FQI learns much more. Known-model planners achieve value with the policy-visible record. Under repetitive logs, deterministic CQL and Decision Transformer copy the behavior policy's most likely action and lose value against that stochastic policy. How can learners use partially observed, observation-driven logs? How can deterministic policies avoid copying repetitive logging habits without losing value? How can long-horizon logs support even simple schedule estimation when importance weights collapse?

Reproducibility statement. The anonymized supplement provides a lockfile, a public API with a seven-line example, POMCP and QMDP planners, tests, aggregate reports, and the release's `B1_REPORT.md` and `FRESH_ENV_RECEIPT.json`. Reproduction commands and recorded targets are included. Author fresh-environment receipts report exact Respiratory values and two scoped full CPU rerun fits, CQL under S1 and DQN/FQI under S0, matched recorded results within stated tolerances; selected CPU tensors differ from the recorded CUDA tensors. Credentialed MIMIC-IV users can rebuild inputs with the included constructor; the authors' raw MIMIC-IV v3.1 rebuild matched the recorded aggregate input identity. Empirical inputs and checkpoints are not redistributed under the data-use agreement. Appendix C documents hyperparameters and seeds; Appendices A and N give commands and source records.

Ethics statement. Synthetic policy values do not establish clinical treatment effects or support clinical decisions.

AI use statement. AI coding and language assistants helped implement benchmark and analysis code, prepare experiment protocols, run and summarize analyses, and edit the manuscript for clarity. The authors specified the research questions and study designs, set each experiment protocol before its outcomes, verified code through tests and source records, checked every reported number against source records, and take full responsibility for all content.

REFERENCES

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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097, 2021.

Kartik Choudhary, Dhawal Gupta, and Philip S. Thomas. ICU-Sepsis: A benchmark MDP built from real medical data. *Reinforcement Learning Journal*, 2024. URL https://rlj.cs.umass.edu/2024/papers/RLJ_RLC_2024_194.pdf.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning, 2020. URL https://arxiv.org/abs/2004.07219. arXiv:2004.07219.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a8166da05c5a094f7dc03724b41886e5-Paper.pdf.

Scott Fujimoto, Edoardo Conti, Mohammad Ghavamzadeh, and Joelle Pineau. Benchmarking batch deep reinforcement learning algorithms, 2019. URL https://arxiv.org/abs/1910.01708. Deep RL Workshop at NeurIPS 2019; arXiv:1910.01708.

Omer Gottesman, Fredrik D. Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christopher Mosch, Li-Wei H. Lehman, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David A. Sontag, and Finale Doshi-Velez. Evaluating reinforcement learning algorithms in observational health settings, 2018. URL https://arxiv.org/abs/1805.12298. arXiv:1805.12298.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019. doi: 10.1038/s41591-018-0310-5.

Mason Hargrave, Alex Spaeth, and Logan Grosenick. EpiCare: A reinforcement learning benchmark for dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130536–130568, 2024. doi: 10.52202/079017-4149.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. doi: 10.1038/s41591-018-0213-5.

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

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020.

Ori Linial, Guy Tennenholtz, and Uri Shalit. Benchmarks for reinforcement learning with biased offline data and imperfect simulators, 2024. URL https://arxiv.org/abs/2407.00806. arXiv:2407.00806.

Zhiyao Luo, Mingcheng Zhu, Fenglin Liu, Jiali Li, Yangchen Pan, Jiandong Zhou, and Tingting Zhu. DTR-Bench: An in silico environment and benchmark platform for reinforcement learning based dynamic treatment regime, 2024. URL https://arxiv.org/abs/2405.18610. arXiv:2405.18610.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. AWAC: Accelerating online reinforcement learning with offline datasets, 2020. URL https://arxiv.org/abs/2006.09359.

Alizée Pace, Hugo Y'eche, Bernhard Schölkopf, Gunnar Rätsch, and Guy Tennenholtz. Delphic offline reinforcement learning under nonidentifiable hidden confounding, 2023.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *International Conference on Learning Representations*, 2025.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pp. 147–163, 2017. URL https://proceedings.mlr.press/v68/raghu17a.html.

David Silver and Joel Veness. Monte-carlo planning in large POMDPs. In *Advances in Neural Information Processing Systems*, volume 23, 2010. URL https://papers.neurips.cc/paper_files/paper/2010/file/edfbe1afcf9246bb0d40eb4d8027d90f-Paper.pdf.

Yihao Sun, Jiaji Zhang, Chengxing Jia, Haoxin Lin, Junyin Ye, and Yang Yu. Model-Bellman inconsistency for model-based offline reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33177–33194, 2023. URL https://proceedings.mlr.press/v202/sun23q.html.

Guy Tennenholtz, Shie Mannor, and Uri Shalit. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B*, 61(3):611–622, 1999. doi: 10.1111/1467-9868.00196.

Yuepeng Wang, Ken Kawano, Yoshihiko Fujisawa, Yongqi Zhou, Akifumi Wachi, Mehrshad Sadria, Lei Zhou, Richard Weiss, Katsuki Fujisawa, Ying Chen, Xin Liu, Kyoung-Sook Kim, Xiao Hu, Sebastien Gros, and Xun Shen. A unified benchmark for dynamic medical treatment reinforcement learning, 2026. URL https://arxiv.org/abs/2606.01028. arXiv:2606.01028v2.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992. doi: 10.1007/BF00992698.

Chuan Wen, Jierui Lin, Trevor Darrell, Dinesh Jayaraman, and Yang Gao. Fighting copycat agents in behavioral cloning from observation histories. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1b113258af3968aaf3969ca67e744ff8-Abstract.html.

Qianyi Xu, Gousia Habib, Feng Wu, Dilruk Perera, and Mengling Feng. MedDreamer: Model-based reinforcement learning with latent imagination on complex EHRs for clinical decision support, 2025. URL https://arxiv.org/abs/2505.19785.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. COMBO: Conservative offline model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/file/f29a179746902e331572c483c45e5086-Paper.pdf.

APPENDIX GUIDE

A. How can a new method be run and evaluated? (p. 12)

B. What are the numerical task and observation laws? (p. 13)

C. How were learners trained and selected? (p. 15)

D. What exact opportunity and controller values are available? (p. 17)

E. What does each information set support under fixed rules? (p. 18)

F. What were the learner scores, full leaderboard, scaling results, and visibility–coverage control? (p. 19)

G. How do scores compare within each timing? (p. 25)

H. Which condition contrasts and components were observed? (p. 25)

I. How were offline actions logged? (p. 26)

J. What do persistence, ablation, reset, and sampling show? (p. 29)

K. Which actions were selected and how often were they logged? (p. 33)

L. What changed under the tested tuning grid? (p. 34)

M. How often did observed source-cohort settings repeat? (p. 35)

N. Where are the exact source records and file identities? (p. 35)

O. How does this case relate to prior benchmarks? (p. 36)

A USING EHRGYM

```
from ehrgym import benchmark as eb
suite = eb.load_suite("respiratory", demo=True)
data = suite.dataset(cell="M1T1S0", replicate=1)
method = eb.baseline("bc")
policy = method.fit(data)
report = suite.evaluate(policy, cell="M1T1S0", episodes=32, replicate=1)
print(report.normalized_score, report.diagnostics)
```

The supplementary release provides a locked CPU environment, exact-reference command, full CQL M1/T1/S1 and DQN/FQI M1/T1/S0 confirmation commands, and a quickstart. Its invented-input demo and scientific-input smoke use reduced budgets and do not reproduce paper results. The code-only archive includes full-record and history-only POMCP/QMDP planners. Reduced-budget controller smoke runs check execution and exact references; full-budget controller commands are documented but were not rerun for the archive. A new method trains on offline training logs, selects on separate logs, uses only public observations at evaluation, and reports independent logged-data

replicates and uncertainty. Report the condition, input identity, method, replicate, selection rule, uncertainty unit, timing-specific NS, raw return, reward components, actions, and aggregate summaries of selected-action behavior-policy propensities (mean, quantiles, and low-support fractions). Compare M1/T1 with clearly labelled controller episode sets. A full submission includes eight-cell returns and action, behavior-policy support, and reward diagnostics; the public API exposes these aggregate propensity summaries under the requested S0/S1 condition, not per-decision selected-action probabilities.

Exact references need no MIMIC input. Scientific learner runs require corrected aggregate inputs, which credentialed users can reconstruct from MIMIC-IV v3.1. The authors' reconstruction reproduced the recorded aggregate input identities; the scoped CPU learner reruns used the existing corrected inputs. These checks do not establish external independent reproduction. The archive contains no empirical inputs or checkpoints; distribution of MIMIC-derived aggregate parameters remains undecided. Commands and exact source files are listed in Appendix N.

B Respiratory Simulator Specification

At each four-hour tick, the finite latent clinical state includes oxygenation burden (three levels), binary recruitability, binary gas-exchange response and vulnerability, a four-level high-FiO_2 exposure run, and three-level pending PEEP, FiO_2, and toxicity phases. The full state also stores the previous one of 12 supported PEEP/FiO_2 actions. The initial burden probabilities are $(0.25, 0.50, 0.25)$; the three binary classes are independently uniform, pending phases and exposure run start at zero, and previous action starts at 0. This is an authored initialization, not an estimated patient distribution.

For each axis, the action and latent context determine a target level and admissible beneficial, worsening, toxicity, or null response routes. The nominal numeric kernel assigns route probabilities by whether the selected level matches, under- or overtreats the target and by response class. It combines the PEEP and FiO_2 route draws, applies immediate or pending effects, updates exposure memory, and draws termination from updated oxygenation burden. Stable burden can terminate in stabilization before the horizon, severe burden in a critical event; moderate burden continues until the last tick, when termination is forced. This is a synthetic mechanism, not a clinical effect estimate.

The observation background is a train-only rank-four aggregate-data fit; latent burden shifts emitted normalized measurements through fixed response directions. M0 uses empirical-aggregate acquisition rates by time stratum. M1 redistributes acquisition with severity, prior availability, recency, and previous action while matching the acquisition load; acquired laboratory results can arrive at the current or next tick. Released observations contain measured values, observed/new masks, recency and availability ages, action and timing fields. Unreleased values and the clinical state remain hidden.

Under T0 the action hold is one tick. Under T1 a one-, two-, or three-tick hold is drawn from the softmax of $(s(2b-1), 0, -s(2b-1))$, where b is oxygenation burden divided by two and $s = 1.386$ is the authored timing-logit scale. Holds are truncated by the remaining 11-tick horizon or earlier terminal event. At each held tick the primary reward is $-0.05I(a)^2$, with $+1$ upon stabilization or -1 upon a critical event, where $I(a)$ is the codec's normalized action intensity. Tick rewards within a hold and across decisions use $\gamma = 0.99$ per physical tick. Acquisition rates and observation background are empirical-aggregate inputs; state classes, response/termination probabilities, measurement adjustments, timing, initialization, and reward are authored. The transition and observation source files are listed in Appendix N.

B.1 Numerical action, transition, and observation laws

Dense action IDs $0, \ldots, 11$ map in order to PEEP/FiO_2 level pairs $(0,1), (0,3), (0,4), (1,1), (1,3), (1,4), (3,1), (3,3), (3,4), (4,1), (4,3), (4,4)$. The corresponding source action ID is $5\,\mathrm{PEEP} + \mathrm{FiO_2}$; source IDs not in this list are unavailable. If $r_P \in \{0,1,2,3\}$ and $r_F \in \{0,1,2\}$ are the positions of the selected levels in their supported lists, normalized intensity is $I(a) = \sqrt{((r_P/3)^2 + (r_F/2)^2)/2}$. Dense action 0 is $(0,1)$ and has zero normalized intensity but remains a supported treatment setting.

The nominal synthetic response law draws independent PEEP and FiO_2 routes conditional on state and action. For burden $b = 0$, the target is $(0,1)$. For $b = 1$, target PEEP is 3 if recruitable and 1

otherwise, while target FiO_2 is 3 for high gas-exchange response and 4 otherwise. For $b = 2$, target PEEP is 4 if recruitable and 3 otherwise, and target FiO_2 is 4. A matched axis has beneficial-route probability 0.65 only when burden is positive and its response class allows that route (recruitable for PEEP, high gas-exchange response for FiO_2); otherwise its beneficial probability is zero. Its background-worsening probability is 0.10. Matched-route toxicity has probability 0.10 only at the maximum axis level with vulnerability, and for FiO_2 only after at least two consecutive high-FiO_2 actions. Undertreatment has worsening probability 0.45; overtreatment has worsening probability 0.20 and, when the adverse route is allowed, toxicity probability 0.25. For overtreatment, PEEP toxicity is allowed at levels 3 and 4 with vulnerability or at level 1 with stable burden and vulnerability; FiO_2 toxicity is allowed at levels 3 and 4 with vulnerability after at least two consecutive high-FiO_2 actions. Undertreatment draws worsening but no toxicity. Remaining mass is the null route. A beneficial route lowers burden one level; worsening raises it one level; toxicity enters the pending toxicity phase. Beneficial effects apply at the current tick with probability 0.70 and at the next tick with probability 0.30, independently by axis. Pending efficacy and toxicity phases apply before the next route draw; toxicity raises burden one level. The high-FiO_2 exposure counter increases by one, capped at three, for selected level 4 and otherwise resets to zero. Before the last tick, burden zero stabilizes with probability 0.12 and burden two has a critical event with probability 0.10; burden one continues. At the final tick, burden zero stabilizes, burden two has a critical event, and burden one splits equally between these outcomes.

The background emission draws 29 direct transformed features from the train-only rank-four aggregate fit, then adds a synthetic state shift. In Respiratory, the shift subtracts $0.10b$ from each of the two oxygenation basis directions and adds $0.10r_F$ to observed FiO_2's transformed coordinate, where r_F is the previous action's FiO_2 position. Each transformed direct value z_j maps to raw units as $c_j + s_j \sinh(z_j)$ using the aggregate-fit center c_j and scale s_j; derived values are recomputed from raw direct values. The empirical centers, scales, loadings, and acquisition rates are credentialed aggregate inputs, so their individual coefficients are not reproduced here.

For direct feature j, let p_j be its empirical M0 acquisition probability in the early, middle, or late time stratum. M0 uses p_j. M1 first sets all feature logits to $\mathrm{logit}(p_j) + 0.3m_j + 0.2\min(\ell_j/12, 1)$, where m_j is prior availability and ℓ_j is recency in hours. The target-group probabilities follow these history-shifted logits; a scalar intercept on nontarget logits preserves the total M0 acquisition load. Subsequent severity and action transfers redistribute that load. Specifically, the M1 repair transfers up to $0.06b$ times the number of respiratory targets from nontarget donors to target recipients and up to $0.03r_P$ times the number of PEEP-focus recipients from respiratory-group donors to SpO_2/PaO_2 recipients, with probabilities bounded away from zero by 10^{-8}. Each transfer is proportional to donor available mass and recipient remaining capacity and conserves total acquisition load. M0 laboratory results arrive at the current tick with probability 0.70; M1 target laboratories use clipped probability $0.55 + 0.30b + 0.10r_P$, and nontarget laboratories use clipped $0.55 + 0.15b$, each clipped to $[0.05, 0.95]$. Bedside features arrive at the current tick. Otherwise an acquired result arrives one tick later. Here b is burden divided by two in acquisition and timing formulas, and r_P is normalized previous-action PEEP position; the transition target rule above uses integer burden.

For exact values, write $K_t(s', d, z, R \mid s, a)$ for the known model's distribution of next decision state, held duration, terminal indicator, and within-hold discounted reward at physical tick t. Backward evaluation uses

$$V_t(s) = \max_a \sum_{s',d,z,R} K_t(s', d, z, R \mid s, a)\{R + \mathbf{1}[z = \text{continue}]\gamma^d V_{t+d}(s')\}, \qquad V_{11} = 0.$$

The hidden-state oracle uses the maximum at every decision; each constant-action value fixes the same a throughout and averages over the stated initial distribution. The best constant is the maximum of those constant-action values. The transition law includes the T0 one-tick schedule or T1 softmax duration rule above, truncated at termination and the horizon. This recurrence evaluates the authored model, not an information-restricted optimum.

B.2 Information and intervention boundaries

The observation process uses aggregate MIMIC-IV version 3.1 measurement rates, corrected source mappings, and a train-only rank-four probabilistic principal-component fit for background variation. Latent response dynamics and treatment effects remain declared synthetic mechanisms. A policy

sees released public observations and its own prior actions, not evaluator hidden state or future results. Decision Transformer additionally uses realized reward after each interval as public feedback for its return target. The action inventory consists of 12 supported PEEP/FiO_2 pairs. The smallest normalized-intensity pair is a supported action and must not be described as an absence of respiratory treatment.

M changes measurement acquisition and availability, with M1 intended as the more informative level. T0 gives a one-tick decision schedule, whereas T1 uses a hidden-burden-dependent one-to-three-tick schedule. Released elapsed duration can signal burden as well as change action opportunities. S0 and S1 create distinct offline training logs. S1 retains positive support but is more concentrated and persistent. Across S0/S1, the evaluator's response, reward, and action laws are unchanged. A support contrast therefore combines these two behavior-policy features. Behavior-policy point comparisons lack paired intervals. The S contrast bundles concentration and persistence. The factorial follow-up tests their specified levels separately under unchanged M1/T1 synthetic laws. No factorial contrast is a causal effect of a real clinical intervention.

B.3 Uncertainty and comparisons

The exact best constant and hidden-state oracle use finite-state evaluation. The full-record and history-only POMCP values, with their earlier QMDP counterparts, are executed policy estimates. Full-record POMCP in the shared-seed run shares 1,024 environment seeds with history-only POMCP, while full-record POMCP in a separate run uses an independent episode set. The two planners use independent filter and planner seeds; actions and simulator random-number consumption can differ. Earlier history-only QMDP and POMCP share a seed namespace and may overlap. The paired full-record versus history-only QMDP audit uses a new, separate episode set of 4,096 episodes for each timing. Their descriptive episode-based intervals do not share the learner-panel uncertainty unit. The 12-method screen has one logged-data replicate and one optimization-seed index. The six-method confirmation has independent logged-data replicates 1–3, optimization-seed index 0 with numerical seeds derived per block, and 8,192 policy-evaluation episodes per fit. Six methods across eight conditions and three new logged-data replicates require $6 \times 8 \times 3 = 144$ confirmation fits. For a method or factor, each replicate contributes one eight-cell or marginal average. The two-sided 95% intervals use $t_{0.975,2} = 4.303$ times the standard error of these three replicate averages. They are nominal and unadjusted. The exact constant is subtracted without sampling error. Logged-behavior differences are descriptive because a paired logged-behavior uncertainty calculation was not retained. No optimization-seed or method-family population uncertainty is estimated. The behavior-policy factorial total-return contrasts pair three data blocks across corners and use Bonferroni 95% family intervals across nine prespecified comparisons; its component and action metrics are descriptive.

C Development Screen and Confirmation Methods

The 12-method screen covers imitation, value learning, support-aware and conservative learning, sequence modeling, and model-based approaches. The screen manifest was frozen before its scientific fits; the final six-method manifest was frozen before the new-data confirmation. They define the implementation and checkpoint-selection rule for every entrant. The benchmark implementations follow a published core objective but make EHRGym-specific adaptations to public histories, discrete legal actions, and physical-time durations. No entry is claimed to reproduce its authors' official code or original benchmark hyperparameters. For BC, DQN/FQI, and CQL, the shared step projection and one-layer GRU have width 128 and use at most 11 public steps. The remaining model-free entrants retain a frozen history encoder; the sequence and model-based rows describe their different architectures below.

Table 5: **Frozen twelve-method EHRGym screen.** Adaptations, readouts, and offline checkpoint metrics are taken from the final method manifest. Every selection metric is minimized on a separate 2,048-episode logged selection set; no evaluation return selects a checkpoint.

Method	Core objective and benchmark adaptation	Action output and checkpoint metric
BC	Masked action imitation on the public-history GRU; no Bellman update.	Samples masked softmax; held-out masked action cross-entropy.
DQN/FQI (Watkins & Dayan, 1992)	Discrete history-Q fit with Huber TD loss, hard target sync every 250 updates, and SMDP target using γ^d.	Masked argmax; held-out Huber TD residual.
BCQ (Fujimoto et al., 2019)	Ported discrete BCQ support-filter objective on the history GRU; action threshold 0.3, logit penalty 0.01, eligible double-Q target with γ^d.	Masked eligible argmax; held-out eligible-target TD residual.
CQL (Kumar et al., 2020)	Frozen discrete CQL penalty $\alpha = 1$ on the history-Q SMDP target.	Masked argmax; held-out Huber TD residual.
IQL (Kostrikov et al., 2022)	Ported expectile-value and advantage-weighted categorical actor objective on history features; expectile 0.7, temperature 3, SMDP value target.	Masked argmax; held-out Q_1 MSE against $r + \gamma^d V(s')$.
AWAC (Nair et al., 2020)	Discrete EHRGym actor with twin policy-evaluation critics and clipped $\exp(A/\lambda)$ weights, $\lambda = 1$; SMDP target.	Masked argmax; held-out policy-evaluation TD MSE.
TD3+BC (Fujimoto & Gu, 2021)	New categorical actor adaptation; expected-value target smoothing replaces Gaussian action noise, with behavior-cloning term, $\alpha = 2.5$, delayed actor update.	Masked argmax; held-out TD MSE against target-actor expectation.
Decision Transformer (Chen et al., 2021)	Ported 3-layer, width-128, one-head transformer with physical-tick embedding and SMDP-discounted return-to-go conditioning.	Masked argmax; held-out action cross-entropy conditioned on logged return-to-go.
MOPO (Yu et al., 2020)	New discrete-SAC EHRGym adaptation on a shared seven-member dynamics ensemble (five elites), five-step model rollouts, uncertainty reward penalty, real-data ratio 0.05.	Masked argmax; held-out real-data soft-TD MSE.
COMBO (Yu et al., 2021)	Same frozen ensemble and rollouts with a model-state conservative value regularizer, coefficient 1, and real-data ratio 0.5.	Masked argmax; held-out real-data soft-TD MSE.
MOBILE (Sun et al., 2023)	Same frozen ensemble and rollouts with Bellman-target disagreement penalty, coefficient 1, and real-data ratio 0.05.	Masked argmax; held-out real-data soft-TD MSE.
MedDreamer (Xu et al., 2025)	Compact AFI and categorical recurrent state-space model with a duration head; two-phase real and hybrid-imagination actor-critic, five-step imagination, SMDP target.	Masked argmax; held-out real-data latent-critic TD MSE.

All methods train on 16,384 logged episodes per cell/replicate and select from 2,048 separate logged episodes. BC and Decision Transformer receive 3,000 updates; most other entries receive 6,000, with MedDreamer's world and policy phases separately budgeted in its frozen configuration. Candidates are assessed every 1,000 updates; ties choose the earliest update and then the checkpoint identifier. The six confirmed methods are BC, DQN/FQI, CQL, IQL, Decision Transformer, and MedDreamer. For Decision Transformer, the deployed initial target is the 90th percentile of `train.initial_rtg`; after each interval, public realized reward r and duration d update it as $G_{\text{next}} = (G_{\text{current}} - r)/\gamma^d$. The wrapper forwards the environment reward after the transition. This is realized feedback, not future information; the other confirmed methods do not use this return-target update. They retain the screen configurations and use independent data replicates 1–3 with optimization-seed index 0 and derived numerical seeds per block. The confirmed optimization settings appear in Table 6; the supplementary method manifest retains the full configurations.

C.1 Confirmed learner settings

The confirmed panel uses the settings below for every cell and data replicate. All use batch size 512 and checkpoint candidates every $1{,}000$ updates. The common model-free learning rate is 0.0003 and weight decay 10^{-5}; Decision Transformer uses learning rate 0.0001 and weight decay 0.0001. Separate held-out logs choose the minimum method-specific metric in Table 5; ties choose the earliest candidate. No direct-evaluation return selects a confirmation checkpoint. MedDreamer additionally uses world-model learning rate 0.0003, phase-one fraction 0.5, advantage-weighting scale 1.0, and weight clip 100; Decision Transformer uses return-to-go scale 1.0 and target quantile 0.9.

D Task Values and Reference Ladder

The exact M1/T1 values appear in Table 7; best-constant to hidden-state-oracle headroom is about 0.6262 return. This gap need not represent value caused by observing physiology: a time-varying policy can outperform a constant without using measurements. The clock-schedule reference below quantifies such time-only value. Full-record QMDP in a separate run reference is -0.3447 over 4,096 episodes, a $+0.2380$ gain. The full-record POMCP in a separate run reference is -0.3242 over 1,024 episodes, a $+0.2585$ gain and a descriptive 41.3% recovery of the exact hidden-state-oracle gap. History-only POMCP reaches -0.3856 over 1,024 episodes, a $+0.1971$ gain and 31.5% point recovery. Full-record POMCP in the shared-seed run returns -0.3614. POMCP uses 16,384 planning simulations per decision and a known synthetic model. Its descriptive gain interval, $[+0.2004, +0.3166]$, is episode based. Both planners are constructive known-model references, not information-set optima or offline learners. Shared-seed run full-record minus history-only POMCP return is $+0.0241$, descriptive paired t 95% interval $[-0.0231, +0.0714]$. The unpaired contrast with the separate run is $+0.0614$, descriptive normal 95% interval $[-0.0196, +0.1424]$. Neither POMCP contrast establishes a physiology-specific benefit; the paired QMDP audit is reported in Table 9.

D.1 Clock-schedule reference

A schedule chooses a ventilator setting only from the physical tick at a decision, without using observations, previous actions, or realized hold durations. Under T1, the chosen action persists through the sampled hold. Coordinate ascent from all constant schedules and width-8 beam search found a T1 schedule with exact expected M1/T1 return -0.421424, or 25.76 NS between the best constant -0.582708 and hidden-state oracle $+0.043489$. It was stable to every single-tick action change. Search did not exhaust all schedules, so its exact value is a lower bound on the best open-loop schedule, not an open-loop optimum. The same method found a T0 schedule returning -0.380210, or 23.89 NS on the separate T0 scale. These values are model expectations and have no sampling interval.

The T1 decision-tick actions for ticks 0–10 are $(7, 10, 7, 7, 10, 8, 7, 7, 10, 8, 10)$; T0 actions are $(7, 1, 10, 8, 7, 1, 10, 8, 10, 8, 7)$. During a T1 hold, the prior action remains active even when a later tick has a different table entry. The T1 schedule was evaluated exactly at -0.421424; a separate 8,192-episode native simulator check returned -0.421642 (SE 0.010086). Because the schedule ignores observations, the check used the invented demo observation decoder with the native clinical transition and reward mechanism. The schedule's NS is computed from the T1 anchors; T0 uses its

Table 6: **Confirmed optimization settings.** Updates count minibatch steps; MedDreamer has separate world and policy phases.

Method	Updates	Additional fixed settings
BC	3,000	Masked softmax sampling
DQN/FQI	6,000	Hard target sync every 250 updates
CQL	6,000	Conservative coefficient $\alpha = 1$
IQL	6,000	Expectile 0.7, temperature 3, weight clip 100, target $\tau = 0.005$
Decision Transformer	3,000	3 layers, 1 head, width 128, dropout 0.1
MedDreamer	$6{,}000 + 6{,}000$	Imagination horizon 5, $\lambda = 0.95$, entropy 0.0003, target $\tau = 0.02$

own hidden-state oracle. History-only POMCP at 31.5 NS and full-record POMCP at 35.3 NS in the shared-seed run, and full-record QMDP at 38.0 NS in a separate run, are executed-policy estimates. Their point differences from the exact schedule do not identify causal information value. Source: `evidence/respiratory_open_loop_schedule_v1/REPORT.md` and `RESULT.json` in the supplementary material.

D.2 Coverage and calibration scope

The S0 behavior policy assigns every supported action at least $0.02/12 = 1/600$ probability through its uniform component; its observation-dependent term can raise that probability. Some S0 actions remain rare: in the M1/T1/S0 DQN/FQI action audit, the fifth percentile of selected-action propensity is 0.0070, and 8.03% of selected decisions have propensity below 0.01. Positive support therefore does not imply adequate coverage of useful action sequences.

Calibration here covers aggregate measurement acquisition rates and a train-only observation background from MIMIC-IV v3.1. Exact constant and hidden-state-oracle values and executed known-model controllers characterize the authored response law. Treatment responses, terminal events, rewards, and the behavior policy were not calibrated to patient outcomes or clinician action probabilities. The known-model controller gains do not establish that a policy can learn the same gains from the finite logs.

D.3 Respiratory planning-budget sensitivity

A prespecified 16-episode diagnostic compared 1,024 and 4,096 planning simulations per decision with 16,384 on matched episodes: lower-minus-retained returns were $+0.2346 \pm 0.2308$ and $+0.1376 \pm 0.2644$ SE, respectively. First-action agreement was 25.0% and 56.25%, respectively; Limitations discusses unresolved planner convergence.

E Restricted-History Reference Controllers

This section compares the history-only planner with the full-record planner under the same planning rule.

E.1 History-only POMCP reference

At Respiratory M1/T1, history-only POMCP receives physical tick, decision index, previous action, elapsed duration, and continuation. It receives neither physiological values nor acquisition metadata, hidden state, realized reward feedback, or future information. The exact 2,592-state filter conditions on duration and survival; the 16,384-simulation planner is approximate. Numerical transition, belief-update, actor-field, and matched-seed invariance checks were performed for this restricted-history reference. Duration and survival can carry severity information, so the policy is neither strictly open-loop nor the optimum among all no-physiology policies. History-only POMCP mean total return is -0.385562 (episode SE 0.028800, descriptive t 95% interval $[-0.442076, -0.329048]$, $n = 1024$). History-only QMDP is a supplementary known-model diagnostic: mean -0.397945 (episode SE 0.014431, descriptive t 95% interval $[-0.426238, -0.369652]$, $n = 4096$). Full-record

Table 7: **Respiratory reference evidence at M1/T1.** The full-record POMCP recovery ratio from the separate run is a descriptive point estimate.

Check	Respiratory evidence
Exact hidden-state opportunity	Best constant -0.5827077790129972; hidden-state oracle $+0.043488768904643274$; headroom 0.6262.
Full-record planner numerical checks	Filtering/source-adapter parity checks; POMCP public-likelihood total-variation certificate below 5.2×10^{-5}.
Executed full-record planner gain	Full-record POMCP in a separate run -0.3242, gain $+0.2585$; full-record POMCP in the shared-seed run -0.3614, gain $+0.2213$.
Constructive headroom recovery	Separate-run full-record POMCP point recovery 41.3%.
Event-component benefit	Full-record POMCP in a separate run versus constant: clinical $+0.3041$, treatment -0.0456; shared-seed run full-record minus history-only POMCP: clinical $+0.0440$ $[-0.0034, +0.0914]$, treatment -0.0198 $[-0.0248, -0.0148]$.

POMCP in the shared-seed run returns -0.361414; its paired difference from history-only POMCP is $+0.024148$ with interval $[-0.023079, +0.071374]$. The history-only POMCP gain is 89.1% of the full-record POMCP gain in the shared-seed run, with delta-method interval $[68.8\%, 109.4\%]$. Full-record POMCP in a separate run returns -0.324185; its unpaired difference from history-only POMCP is $+0.061376$ $[-0.019632, +0.142385]$ and the separate-run history-only/full-record POMCP point ratio is 76.3%. These are distinct episode-set estimates, not a pooled result or physiology decomposition. The paired estimates use the retained planner evaluations. The full-record planner's two evaluation episode sets differ by about 0.037 in mean return, beyond the paired difference; episode intervals omit independent planner-seed variation.

E.2 Paired QMDP information comparison

The full-record QMDP planner uses the public EHR packet; the history-only QMDP planner uses actions, timing, and continuation history. Each timing uses its own exact hidden-state Q table and a new paired episode set of 4,096 environment seeds; T1 is primary and T0 secondary. Earlier full-record QMDP estimates from a separate run suggested the T1 question before this prospective episode set was evaluated. Both prespecified timing contrasts remain significant after Holm correction ($p = 9.434 \times 10^{-5}$ each). At T1 the total contrast is $+8.0$ NS points; T0 has a different oracle denominator, so its normalized contrast must use T0 anchors. These executed planners are not belief-space POMDP optima, and the contrast does not isolate causal mediation by individual measurements or replace the inconclusive paired POMCP comparison. The comparison used a prespecified analysis plan (in the supplement) and a new evaluation episode set.

Planning-budget diagnostics found different selected actions at two of eight preselected roots between 4,096 and 16,384 simulations, and independent full-budget seeds differed at two of four later histories. These checks do not establish planner convergence. The restricted actor also excludes reward feedback, so this comparison cannot isolate physiology from every other possible policy information set.

E.3 Timing-law control

Only the authored T1 hold-logit scale changes: the original 1.3863, half scale, or zero. At zero scale, raw hold durations are independent of hidden burden; continuation and horizon truncation can still carry state information. Exact anchors are recomputed for each scale. Full-record and history-only QMDP share 4,096 episode seeds per scale. This timing-law control evaluates QMDP only; POMCP was not run.

F Development Screen and Confirmation Results

Table 11 reports the 12-method screen once alongside confirmation and known-model references. Development and confirmation use separate data replicates; planner episode sets are labelled. The

Table 8: **Shared-seed run full-record minus history-only POMCP contrast.** Stabilization, critical-event, and treatment rows are signed, discounted return components. Descriptive paired episode t intervals use 1,024 matched environment seeds; ratio interval uses the delta method. Full-record POMCP in a separate run uses an independent episode set.

Quantity	Point estimate	95% interval
Stabilization	$+0.0218$	$[-0.0018, +0.0454]$
Critical event	$+0.0222$	$[-0.0016, +0.0460]$
Clinical sum	$+0.0440$	$[-0.0034, +0.0914]$
Treatment	-0.0198	$[-0.0248, -0.0148]$
Total return	$+0.0241$	$[-0.0231, +0.0714]$
History-only/full-record POMCP gain ratio	89.1%	$[68.8\%, 109.4\%]$
Action-0 frequency: history-only POMCP	0.86%	decision weighted
Action-0 frequency: full-record POMCP (separate run)	0.07%	separate run
Action-0 frequency: full-record POMCP (shared-seed run)	0.04%	shared-seed run

Table 9: **Full-record minus history-only QMDP return.** The paired QMDP run gives the -0.3441 T1 full-record QMDP value; the leaderboard separate run gives -0.3447. Differences use unrounded paired episodes and descriptive t 95% intervals; displayed means are rounded. Stabilization, critical-event, and treatment rows are signed, discounted return components. Within each timing, both planners use the same QMDP rule; T0 and T1 returns are not directly comparable.

Quantity	T1 (primary)	T0 (secondary)
History-only QMDP total return	-0.3941	-0.3935
Full-record QMDP total return	-0.3441	-0.3355
Paired total difference	$+0.0500$ $[+0.0259, +0.0741]$	$+0.0581$ $[+0.0301, +0.0860]$
Stabilization difference	$+0.0228$ $[+0.0108, +0.0348]$	$+0.0227$ $[+0.0088, +0.0366]$
Critical-event difference	$+0.0233$ $[+0.0113, +0.0353]$	$+0.0226$ $[+0.0087, +0.0365]$
Treatment difference	$+0.0040$ $[+0.0015, +0.0064]$	$+0.0128$ $[+0.0096, +0.0160]$

S0/S1 behavior-policy rows each use 8,192 direct policy-evaluation episodes; their descriptive normal 95% intervals equal the recorded mean plus or minus 1.96 episode standard errors and are not paired learner contrasts. All eight cells have timing-specific normalized scores in Table 19.

The development screen's MedDreamer mean is slightly above the exact constant; its independent confirmation mean is below it and has an interval spanning the constant. The 48 screen-to-confirmation cell point signs agree in 44 cells. All four sign disagreements are MedDreamer cells. The mean absolute cell shift is 0.0484 for MedDreamer and 0.0394 for DQN/FQI, versus 0.0040–0.0057 for the other four confirmed methods. This is descriptive stability accounting, not an inferential comparison of the two evidence roles.

The primary M1/T1 numerical estimates appear in Table 12.

F.1 Simple fitted-Q learner ladder at M1/T1/S0

Simple fitted-Q learners test whether the low partial-record score depends on neural optimization. Five independent S0 behavior-policy logs contain 16,384 training episodes each. The tick-only and tick-plus-hidden-state learners use tabular cell means; the tick-plus-exact-belief learner uses unscaled posterior features and ridge least squares. All use the logged held-tick discount and terminal mask. The state learner sees the hidden clinical core during fitting and deployment, making it an information upper control. Iteration count and belief ridge were chosen on 512 separate simulator episodes per candidate and replicate, an optimistic choice with simulator access; untuned defaults are shown alongside. Reporting uses 8,192 disjoint evaluation episodes per fit, never the tuning episodes.

The tick-only learners produce physical-time clock schedules. The table's 1.37 NS is their simulated policy-evaluation mean; exact evaluation of those same five learned schedules gives 1.57 NS $[-4.23, 7.37]$. The separately best-found clock schedule scores 25.76 NS and is a search lower bound, not a proven clock optimum. Its scheduled action was logged at every tick in every replicate. This per-tick action coverage does not establish coverage of action sequences or hidden-state pairs: the

Table 10: **Timing-law QMDP control at M1/T1.** History-only and full-record QMDP entries give return (NS under that scale's anchors, 0–100). Shares are history-only QMDP gains divided by full-record QMDP gains over the scale-specific exact constant; brackets give paired episode 95% t intervals. The original-scale QMDP run has its own paired episode set.

Scale	Constant	Oracle	Share
Zero	−0.5827	+0.0150	0.889
Half	−0.5827	+0.0293	0.748
Original	−0.5827	+0.0435	0.808

Scale	History-only return (NS)	Full-record return (NS)	Measurement record Δ [95%]
Zero	−0.4223 (26.8)	−0.4022 (30.2)	+0.0201 [−0.0047, +0.0448]
Half	−0.4430 (22.8)	−0.3960 (30.5)	+0.0470 [+0.0215, +0.0726]
Original	−0.3902 (30.7)	−0.3445 (38.0)	+0.0457 [+0.0207, +0.0708]

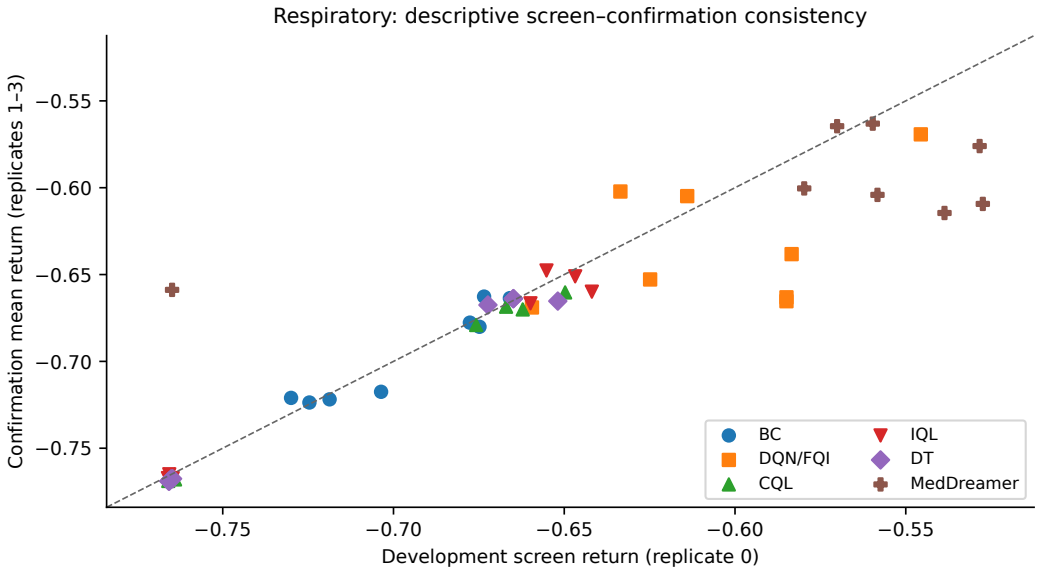

Figure 5: **Development-to-confirmation point consistency in Respiratory.** Each point pairs one method and one M/T/S cell. The horizontal coordinate is the development-screen mean from replicate 0; the vertical coordinate is the independent confirmation mean from replicates 1–3. The dashed diagonal denotes equal point values. The plot is descriptive and carries no paired interval or hypothesis test.

logs visit a mean 0.0609 of possible tick–clinical-core-state pairs. The time step alone is not a Markov state, so Bellman backups over it can fail even if the logs identify a useful clock schedule. The behavior policy chooses actions from observed measurements and the previous action, all available to full-record learners. The tick-only learner cannot see those measurements, so confounding in its logged action-outcome pairs is consistent with its shortfall, not identified as its cause. Two operational attempts stopped after concurrent progress writers collided on a temporary filename; corrected restarts used the same predeclared seeds and unchanged protocol, and failed attempts contributed no estimates. Supplement: `evidence/respiratory_simple_learner_ladder_v1/REPORT.md` and `evidence/respiratory_simple_learner_ladder_v1/RESULT.json`.

F.2 Logged clock-schedule search at nominal M1/T1/S0

Five independent S0 log replicates each supplied 16,384 training episodes and 2,048 separate selection episodes. Schedules were chosen from training logs alone, then evaluated exactly in the synthetic simulator. The first estimator selects at each physical time step the action with the largest observed

Table 11: **Full Respiratory M1/T1 leaderboard.** Confirmation scores use nominal 95% intervals across three independent logged-data replicates; Development has one replicate and no replicate interval; its eight-cell mean and delta subtract the exact best constant. Controllers use descriptive episode intervals, and exact references have no sampling error. Eight-cell learner summaries here retain raw returns; Table 19 supplies timing-specific normalized scores for every cell.

Independent confirmation: six learners, three data replicates

Method	M1/T1 S0 NS [95% CI]	M1/T1 S1 NS [95% CI]	Eight-cell return [95% CI]
BC	-15.2 $[-15.5, -14.9]$	-21.5 $[-25.8, -17.3]$	-0.6960 $[-0.7049, -0.6871]$
DQN/FQI	2.1 $[-2.1, 6.4]$	-13.2 $[-22.4, -4.0]$	-0.6331 $[-0.6554, -0.6108]$
CQL	-13.7 $[-16.1, -11.3]$	-29.3 $[-30.0, -28.7]$	-0.7186 $[-0.7295, -0.7077]$
IQL	-10.4 $[-13.3, -7.4]$	-29.1 $[-30.5, -27.7]$	-0.7116 $[-0.7324, -0.6909]$
DT	-13.0 $[-16.0, -9.9]$	-29.5 $[-29.9, -29.1]$	-0.7167 $[-0.7310, -0.7024]$
MedDreamer	3.1 $[-19.4, 25.7]$	-5.1 $[-23.9, 13.8]$	-0.5989 $[-0.6653, -0.5324]$

Development screen: 12 methods, single replicate, descriptive only

Method	M1/T1 S0 Return	M1/T1 S0 NS	M1/T1 S1 Return	M1/T1 S1 NS	Eight-cell return	Eight-cell Δ constant
BC	-0.6776	-15.2	-0.7036	-19.3	-0.6961	-0.1134
DQN/FQI	-0.5456	5.9	-0.5849	-0.4	-0.6038	-0.0211
BCQ	-0.5944	-1.9	-0.7203	-22.0	-0.6747	-0.0920
CQL	-0.6670	-13.5	-0.7656	-29.2	-0.7144	-0.1317
IQL	-0.6552	-11.6	-0.7657	-29.2	-0.7082	-0.1255
AWAC	-0.6634	-12.9	-0.7661	-29.3	-0.7091	-0.1263
TD3+BC	-0.5758	1.1	-0.6977	-18.4	-0.6353	-0.0526
DT	-0.6649	-13.1	-0.7651	-29.1	-0.7144	-0.1317
MOPO	-0.7148	-21.1	-0.7317	-23.8	-0.7021	-0.1194
COMBO	-0.6513	-11.0	-0.7031	-19.2	-0.6810	-0.0983
MOBILE	-0.7150	-21.1	-0.7138	-20.9	-0.7035	-0.1208
MedDreamer	-0.5597	3.7	-0.5387	7.0	-0.5784	$+0.0043$

Reference rows: exact expectations and known-model controllers

Reference / evidence / episode set	Return [95% CI]	NS [95% CI]
S0 behavior policy		
Direct; descriptive normal, 8,192	-0.6776 $[-0.6943, -0.6610]$	-15.2 $[-17.8, -12.5]$
S1 behavior policy		
Direct; descriptive normal, 8,192	-0.7129 $[-0.7279, -0.6980]$	-20.8 $[-23.2, -18.4]$
Exact best constant		
Exact; no sampling error	-0.5827	0.0
Exact hidden-state oracle		
Exact; evaluator state	0.0435	100.0
History-only POMCP (shared-seed run)		
Episode t, 1,024	-0.3856 $[-0.4421, -0.3290]$	31.5 $[22.5, 40.5]$
Full-record POMCP (shared-seed run)		
Episode t, 1,024	-0.3614 $[-0.4189, -0.3039]$	35.3 $[26.2, 44.5]$
Full-record POMCP (separate run)		
Episode normal, 1,024	-0.3242 $[-0.3823, -0.2661]$	41.3 $[32.0, 50.6]$
Full-record QMDP (separate run)		
Episode normal, 4,096	-0.3447 $[-0.3735, -0.3159]$	38.0 $[33.4, 42.6]$

mean discounted return-to-go among episodes taking that action. It does not adjust for the behavior policy, which chooses actions using observed measurements and the previous action. The second searches schedules by maximizing weighted per-decision importance sampling (WPDIS), using recorded behavior-policy action probabilities. Its cumulative target-to-behavior weights update at decisions and carry through held actions and termination. The selection logs provide estimator diagnostics, not schedule selection.

The per-time-step estimate reaches about half the separately best-found clock schedule's exact 25.76 NS; confounding by observation-driven behavior-policy choices is consistent with this shortfall, not established as its cause. WPDIS search overfits: its mean training estimate is 1.12 return, versus -0.12 on the separate selection logs. Its mean final training ESS is only 1.51 of 16,384 episodes across the 11-step horizon. These different failure modes show that this logged-data

Table 12: **M1/T1 Respiratory learner confirmation.** Entries are mean return [nominal 95% t interval] over three independent logged-data replicates, with optimization-seed index 0 and derived numerical seeds per block. The exact best constant is -0.5827. Figure 1 plots the shared-seed run POMCP pair and the separate run full-record QMDP point; the paired QMDP information comparison and separate run POMCP estimate are reported separately.

Method	S0	S1
BC	$-0.6777\,[-0.6797, -0.6757]$	$-0.7175\,[-0.7441, -0.6909]$
DQN/FQI	$-0.5693\,[-0.5961, -0.5424]$	$-0.6654\,[-0.7229, -0.6078]$
CQL	$-0.6685\,[-0.6833, -0.6537]$	$-0.7664\,[-0.7704, -0.7624]$
IQL	$-0.6477\,[-0.6662, -0.6291]$	$-0.7650\,[-0.7738, -0.7563]$
Decision Transformer	$-0.6638\,[-0.6830, -0.6446]$	$-0.7677\,[-0.7702, -0.7651]$
MedDreamer	$-0.5631\,[-0.7042, -0.4221]$	$-0.6146\,[-0.7326, -0.4965]$

Table 13: **Simple fitted-Q learners on Respiratory M1/T1/S0 logs.** Entries are mean normalized score (NS) [two-sided 95% Student t interval] across five independent log/fit replicates; intervals reflect replicate variation, not evaluation-episode variation. Simulator tuning is an optimistic selection procedure.

Learner input	Simulator-tuned NS [95% t]	Untuned default NS [95% t]
Tick only, tabular	$1.37\,[-3.99, 6.74]$	$1.37\,[-3.99, 6.74]$
Tick and exact belief, linear	$9.43\,[5.20, 13.67]$	$0.99\,[-12.83, 14.81]$
Tick and hidden state, tabular	$41.86\,[39.11, 44.61]$	$41.44\,[37.99, 44.88]$

schedule search does not close the clock gap; they do not identify the best clock-only schedule. Supplement: `evidence/respiratory_logged_clock_search_v1/REPORT.md` and `evidence/respiratory_logged_clock_search_v1/RESULT.json`.

F.3 Prespecified logged-data scaling at nominal M1/T1/S0

The logged-data scaling study followed a prespecified analysis plan (in the supplement). DQN/FQI, CQL, and MedDreamer each used three new S0 logged-data replicates at each training size, with role-separated selection logs one-eighth as large and 8,192 direct-evaluation episodes per fit (27 fits). Replicate labels and evaluation seeds were paired across sizes within method. Frozen configurations and optimization budgets were held fixed: 6,000 minibatch updates for DQN/FQI and CQL; 6,000 world-model plus 6,000 policy updates for MedDreamer. Thus this varies available logged data, not training duration or model access. Further gains remain possible.

Training sizes are 16,384, 65,536, and 262,144 episodes; selection sizes are 2,048, 8,192, and 32,768. No method has a detectable directional improvement under the fixed optimization budget, although MedDreamer's wide interval includes large gains. Known-model history controllers are references with model access, not trained comparators. The contrast does not identify why learners remain below them, establish zero data effect with longer optimization, or estimate a clinical effect.

F.4 Matched visibility and logging-coverage control

At Respiratory M1/T1, a 2×2 DQN/FQI control crosses observation visibility ($V0$: original available record; $V1$: that record plus the full clinical hidden state) with the behavior policy generating logs ($L0$: original S0; $L1$: uniform over 12 actions). Clinical dynamics, reward, confirmed DQN/FQI configuration, and offline checkpoint selection stay fixed. Each cell uses three new independent log/fit replicates with 16,384 training and 2,048 selection episodes, one matched budget of 18,000 minibatch updates per fit, and 8,192 evaluation episodes per selected checkpoint. NS uses the same exact best constant setting and hidden-state oracle. Table 16 reports means and two-sided 95% Student t intervals across the three log/fit replicates, not across evaluation episodes.

The separate 1,024-episode checkpoint curves in Table 17 describe training, not policy selection. Every cell's mean peaks by 3,000 updates and falls thereafter. Offline selection, using its own logs, chose 1,000 or 2,000 updates in all 12 fits; the selected-score intervals above use the independent

Table 14: **Logged clock-schedule search at M1/T1/S0.** Exact NS is mean [two-sided 95% Student t interval] over five independent log replicates; exact evaluation has no episode-sampling error. Final training effective sample size (ESS) is based on cumulative weights through physical time step 10, out of 16,384 training episodes per replicate.

Schedule selection	Exact NS [95% t]	Final training ESS
Per-time-step logged means	13.24 $[6.34, 20.14]$	—
WPDIS search	-20.85 $[-30.40, -11.30]$	1.51 $[0.92, 2.10]$

Table 15: **M1/T1/S0 logged-data scaling.** Mean NS uses the fixed exact best constant and hidden-state-oracle anchors. Largest-minus-smallest contrasts pair three logged-data replicates; brackets are two-sided 95% Student t intervals ($df = 2$). Holm p values cover the three primary return tests. The middle size is descriptive; intervals are wide, especially for MedDreamer.

Method	16,384 NS	65,536 NS	262,144 NS
DQN/FQI	-2.7	5.2	1.2
CQL	-11.9	-11.5	-13.5
MedDreamer	3.6	0.9	8.6

Method	Return change [95%]	NS change [95%]	Holm p
DQN/FQI	$+0.0247$ $[-0.0103, +0.0596]$	$+3.9$ $[-1.6, 9.5]$	0.28
CQL	-0.0103 $[-0.0363, +0.0156]$	-1.7 $[-5.8, 2.5]$	0.46
MedDreamer	$+0.0311$ $[-0.0705, +0.1327]$	$+5.0$ $[-11.3, 21.2]$	0.46

evaluation episodes. Choosing each fit's best checkpoint on the curve episodes themselves, an optimistic descriptive diagnostic, gives 3.36–6.46 NS for $V0L0$, 8.29–12.58 for $V0L1$, 58.43–63.99 for $V1L0$, and 58.13–73.13 for $V1L1$ (supplementary `evidence/respiratory_visibility_coverage_factorial_v1/REPORT.md`).

F.5 Exact history-belief input control

In a separate Respiratory M1/T1 synthetic simulator control, DQN/FQI received either the original 181-component available record or that record plus an exact 2,592-component posterior over clinical core states. The posterior was computed by the unchanged history-only filter from prior actions, elapsed hold durations, and continuation: the same history available to the history-only planner. Filtering used float64 arithmetic and supplied float32 learner features. No hidden state or future information was supplied to the learner. Both arms used three new matched S0 logged-data and fit replicates, each with 16,384 training and 2,048 offline-selection episodes. Each fit ran 18,000 updates; selected checkpoints used 8,192 fresh evaluation episodes. Within each replicate, arms shared logged data and evaluation seeds. All selected checkpoints were at 2,000 updates.

Across the three paired log/fit replicates, the original-record mean was 0.01 NS (95% Student t interval $[-16.40, 16.43]$); exact-belief input reached 5.05 NS $[-10.93, 21.03]$. The paired belief-minus-original difference was $+5.04$ NS $[2.01, 8.07]$. Intervals use independent paired log/fit replicates as their unit, not evaluation episodes. The prior history-only POMCP 31.5 NS and state-visible S0 DQN/FQI 57.4 NS are descriptive references from different policy and evaluation runs. Separate checkpoint curves did not enter selection; the original-record and belief-arm means at 18,000 updates were -22.65 and -26.82 NS, respectively. This control shows a small paired gain from exact history belief within this DQN/FQI and S0-log design. Source: `evidence/respiratory_belief_input_control_v1/REPORT.md` and `RESULT.json` in the supplementary material.

F.6 Clinical-state-visible, uniform-log learnability control

This Respiratory M1/T1 variant exposes the clinical hidden state in the observation while keeping clinical dynamics unchanged. The exact best constant (-0.582708) and hidden-state oracle (0.043489) were recomputed and unchanged. Its behavior policy samples uniformly across all 12 actions. For

Table 16: **Matched visibility and logging coverage at M1/T1.** Selected DQN/FQI NS and contrasts, each with a nominal 95% replicate t interval. Visibility contrasts average $V1 - V0$ across behavior policies; logging contrasts average $L1 - L0$ across visibility levels; interaction is the V effect under $L1$ minus that under $L0$.

Record and behavior-policy logs	Selected NS [95% interval]
$V0L0$: original record, S0 logs	-0.77 $[-6.22, 4.69]$
$V0L1$: original record, uniform logs	1.55 $[0.85, 2.26]$
$V1L0$: full hidden state, S0 logs	57.36 $[51.77, 62.94]$
$V1L1$: full hidden state, uniform logs	44.59 $[34.54, 54.65]$
Visibility (V)	$+50.58$ $[43.58, 57.59]$
Uniform logging (L)	-5.22 $[-7.38, -3.06]$
Interaction ($V \times L$)	-15.08 $[-26.99, -3.17]$

Table 17: **Descriptive DQN/FQI checkpoint curves.** Mean NS over three fits at each update, using a separate 1,024-episode evaluation set per checkpoint; these episodes did not select checkpoints.

Updates	$V0L0$	$V0L1$	$V1L0$	$V1L1$
1,000	2.4	3.6	57.4	45.6
2,000	0.1	6.7	61.5	62.0
3,000	0.6	6.1	49.9	62.1
18,000	-24.2	-25.0	-18.4	-15.1

each of three independent replicates, learners used 16,384 training and 2,048 offline selection episodes. The six learner configurations and offline checkpoint-selection rule were unchanged except for a threefold policy-update budget; MedDreamer's world-model stage was unchanged. Each selected fit used 8,192 new evaluation episodes. Table 18 gives NS and two-sided 95% Student t intervals across the three independent fits per learner.

The offline rule selected the first, 1,000-update checkpoint in 15 of 18 fits. The low BC and Decision Transformer scores are consistent with imitation of the uniform behavior policy; CQL's score is consistent with conservatism pulling toward that policy. This control does not isolate those mechanisms or show performance under the original partial observation and behavior-policy conditions.

G Normalized Scores in All Eight Respiratory Cells

All eight cells use the same exact best constant, -0.582708, while the exact hidden-state oracle is $+0.264875$ for T0 and $+0.043489$ for T1. Thus the NS denominators are 0.847583 and 0.626197, respectively; neither M nor S changes these anchors under the frozen synthetic law. Each cell below transforms its three-replicate mean and nominal t interval by its own timing denominator. Cross-timing NS values measure fractions of different oracle headroom and cannot be read as a common raw-return scale.

H Cell-Level Outcomes and Factor Contrasts

The following factor contrasts use level 1 minus level 0 and average the other two factors within each logged-data replicate. The main text reports M1/T1 cell values; Table 11 reports the eight-cell raw-return averages. Complete cell means, nominal intervals, and reward components remain in the supplementary aggregate tables; source identities are listed in Appendix N. The behavior policy (S) has the largest observed axis effect. Across eight conditions and three logged-data replicates, the M, T, and S effects are about $+0.01$, $+0.01$, and -0.07 return. S intervals are below zero for BC, CQL, IQL, and Decision Transformer; DQN/FQI and MedDreamer intervals include or touch zero (Figure 6 and Table 21).

At M1/T1/S0, CQL's stabilization, critical-event, and treatment components are $+0.1999$, -0.7379, and -0.1306. At M1/T1/S1 they are $+0.0881$, -0.8538, and -0.0008. The simulated event-

Table 18: **Clinical-state-visible, uniform-log Respiratory M1/T1 control.** NS uses the recomputed exact constant and hidden-state oracle. Intervals are across three independent log/fit replicates, not evaluation episodes.

Learner	Mean NS [95% t interval]
DQN/FQI	$48.0\ [26.0, 69.9]$
IQL	$19.3\ [6.6, 31.9]$
CQL	$5.3\ [-5.0, 15.7]$
MedDreamer	$1.6\ [-13.1, 16.3]$
BC	$-20.9\ [-22.5, -19.2]$
Decision Transformer	$-29.1\ [-43.8, -14.5]$

Table 19: **Timing-specific NS for confirmed learner cells: BC, DQN/FQI, and CQL.** Entries are mean [nominal 95% t interval] across three independent logged-data replicates, with one optimization-seed index and 8,192 evaluation episodes per fit. T0 and T1 use different oracle denominators; negative scores are not clipped.

Cell	BC	DQN/FQI	CQL
M0/T0/S0	$-9.4\ [-10.7, -8.2]$	$-8.3\ [-31.3, 14.8]$	$-11.4\ [-12.4, -10.4]$
M0/T0/S1	$-16.6\ [-20.2, -13.1]$	$-10.2\ [-21.1, 0.8]$	$-21.9\ [-24.2, -19.5]$
M0/T1/S0	$-15.5\ [-16.4, -14.7]$	$-3.1\ [-16.1, 9.9]$	$-12.4\ [-14.7, -10.1]$
M0/T1/S1	$-22.2\ [-24.4, -20.0]$	$-8.9\ [-18.4, 0.7]$	$-29.5\ [-29.7, -29.3]$
M1/T0/S0	$-9.5\ [-10.2, -8.9]$	$-2.6\ [-6.4, 1.2]$	$-10.3\ [-14.1, -6.5]$
M1/T0/S1	$-16.3\ [-18.7, -13.9]$	$-9.5\ [-10.5, -8.5]$	$-22.0\ [-24.0, -19.9]$
M1/T1/S0	$-15.2\ [-15.5, -14.9]$	$2.1\ [-2.1, 6.4]$	$-13.7\ [-16.1, -11.3]$
M1/T1/S1	$-21.5\ [-25.8, -17.3]$	$-13.2\ [-22.4, -4.0]$	$-29.3\ [-30.0, -28.7]$

component sum falls by about 0.228 while treatment penalty falls by about 0.130. These signed components establish the source of the *return arithmetic*; they do not prove a causal mediation path through any one executed action.

I Behavior-Policy Specification

The behavior policy uses only public observation values and the previous action; it is an authored heuristic, not a fit to MIMIC action frequencies. It forms separate PEEP and FiO_2 targets from observed features, skipping unobserved features, clipping standardized values to $[-3, 3]$, and decaying contributions with a 12-hour recency half-life. A sigmoid with scale 1.25 turns each weighted sum into a target between 0 and 1, blended toward 0.5 when few features are available. The PEEP coefficients on public feature values are -0.8 for SpO_2, -0.6 for PaO_2, -0.5 for PaO_2/FiO_2, $+0.3$ for respiratory rate, and $+0.2$ for mean blood pressure. The FiO_2 coefficients are -1.0 for SpO_2, -0.7 for PaO_2, -0.5 for PaO_2/FiO_2, and $+0.3$ for observed FiO_2. These are standardized public features, not raw bedside units. For each axis, action-level logit is $-6(\text{level position} - \text{target})^2$ plus a 0.15 bonus at the previous level; softmax probabilities are mixed with 10% uniform mass. The two axis distributions multiply to form a 12-action base distribution $b(a \mid o, p)$.

Temperature and persistence then give $q_\tau(a \mid o, p) = b(a \mid o, p)^{1/\tau} / \sum_{a'} b(a' \mid o, p)^{1/\tau}$ and $\pi_S(a \mid o, p) = 0.98[(1-\kappa)q_\tau(a \mid o, p) + \kappa \mathbf{1}\{a = p\}] + 0.02/12$. S0 uses $(\tau, \kappa) = (1.0, 0.10)$ and S1 uses $(0.35, 0.75)$. The 0.02 is total uniform mass, giving each action at least $1/600$ from this term. The previous-level axis bonus differs from joint previous-action persistence. Neither hidden state nor reward enters the behavior-policy probability.

Table 20: **Timing-specific NS for confirmed learner cells: IQL, Decision Transformer, and MedDreamer.** Entries are mean [nominal 95% t interval] across three independent logged-data replicates, with one optimization-seed index and 8,192 evaluation episodes per fit. T0 and T1 use different oracle denominators; negative scores are not clipped.

Cell	IQL	Decision Transformer	MedDreamer
M0/T0/S0	-9.9 $[-15.4, -4.4]$	-10.0 $[-11.9, -8.1]$	-3.1 $[-17.7, 11.5]$
M0/T0/S1	-21.8 $[-24.3, -19.2]$	-22.0 $[-24.1, -19.8]$	-9.0 $[-19.9, 1.9]$
M0/T1/S0	-10.9 $[-14.1, -7.7]$	-13.2 $[-16.2, -10.2]$	-2.8 $[-16.0, 10.3]$
M0/T1/S1	-29.5 $[-29.7, -29.3]$	-29.5 $[-30.0, -29.0]$	1.1 $[-9.1, 11.2]$
M1/T0/S0	-9.1 $[-14.6, -3.6]$	-9.6 $[-17.6, -1.6]$	-2.5 $[-10.6, 5.6]$
M1/T0/S1	-21.9 $[-24.0, -19.9]$	-22.0 $[-24.0, -19.9]$	2.1 $[-0.7, 5.0]$
M1/T1/S0	-10.4 $[-13.3, -7.4]$	-13.0 $[-16.0, -9.9]$	3.1 $[-19.4, 25.7]$
M1/T1/S1	-29.1 $[-30.5, -27.7]$	-29.5 $[-29.9, -29.1]$	-5.1 $[-23.9, 13.8]$

Table 21: **Respiratory M/T/S confirmation contrasts.** Values are mean [nominal 95% t interval], based on three independent data-replicate blocks. The fixed six-method panel is not a population sample of method families.

Method	M1–M0	T1–T0	S1–S0
	$+0.0021$	-0.0065	-0.0500
BC	$[-0.0018, +0.0060]$ $+0.0149$	$[-0.0160, +0.0030]$ $+0.0287$	$[-0.0720, -0.0280]$ -0.0516
DQN/FQI	$[-0.0611, +0.0909]$ $+0.0002$	$[-0.0334, +0.0908]$ $+0.0058$	$[-0.1035, +0.0003]$ -0.0982
CQL	$[-0.0072, +0.0077]$ $+0.0027$	$[-0.0075, +0.0191]$ $+0.0078$	$[-0.1032, -0.0932]$ -0.1107
IQL	$[+0.0011, +0.0044]$ $+0.0013$	$[-0.0181, +0.0337]$ $+0.0013$	$[-0.1315, -0.0899]$ -0.1032
Decision Transformer	$[-0.0041, +0.0066]$ $+0.0245$	$[-0.0288, +0.0314]$ $+0.0207$	$[-0.1141, -0.0922]$ -0.0093
MedDreamer	$[+0.0019, +0.0472]$	$[-0.0084, +0.0498]$	$[-0.0577, +0.0392]$

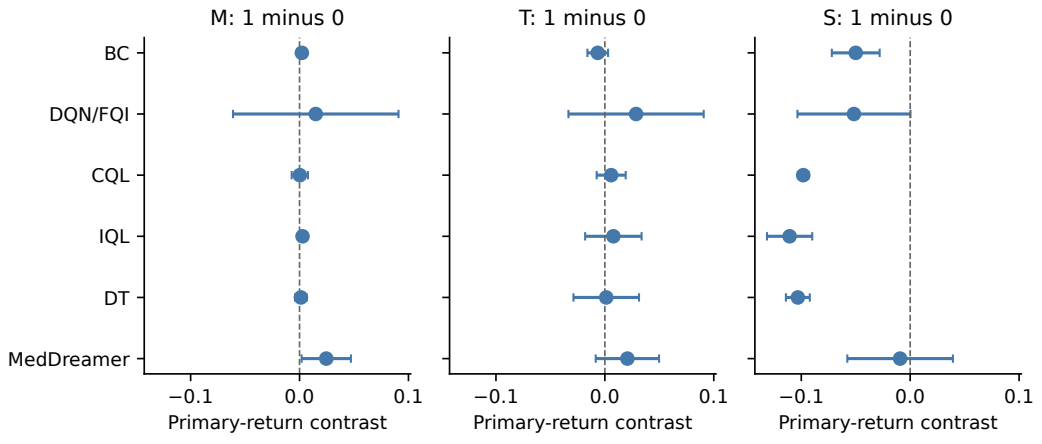

Figure 6: **Respiratory independent-confirmation marginal contrasts.** Each M1–M0, T1–T0, or S1–S0 contrast averages the other two factors within each of three logged-data replicate blocks. Bars are nominal two-sided 95% t intervals with two degrees of freedom. The six methods form a fixed panel. S1 jointly increases behavior-policy concentration and action persistence; no interval compares S's magnitude directly with M or T.

Table 22: **Eight-cell mean reward components in confirmation.** Stabilization, critical-event, and treatment columns are signed, discounted return components; their sum is the simulated total. A less negative treatment value means less treatment penalty. These are confirmed point estimates; component replicate intervals are in the supplementary aggregate tables.

Method	Stabilization	Critical event	Treatment	Total
Exact best constant	$+0.2391$	-0.7002	-0.1217	-0.5827
BC	$+0.1765$	-0.7625	-0.1100	-0.6960
DQN/FQI	$+0.2301$	-0.7072	-0.1560	-0.6331
CQL	$+0.1437$	-0.7961	-0.0662	-0.7186
IQL	$+0.1462$	-0.7938	-0.0639	-0.7116
Decision Transformer	$+0.1447$	-0.7952	-0.0663	-0.7167
MedDreamer	$+0.2563$	-0.6814	-0.1737	-0.5989

J S1 Repetition Controls

J.1 Persistence law

At $\kappa = 0.75$, $\pi(p) \geq (1 - 0.02) \times 0.75 + 0.02/12 = 0.736666\ldots$ (approximately 0.7367) and $\pi(a \neq p) \leq 0.2467$ for every observation; the previous action is therefore the unique conditional mode at both tested concentrations. Starting from reset $p = 0$, the greedy behavior policy selects action 0 throughout, with exact constant value -0.766909. At $\kappa = 0.10$ this conclusion does not hold. This structural comparator alone does not establish learner copying; the separate prespecified reset control tests it (Appendix J.5). Changing reset previous action also changes the lifted initial state.

J.2 Persistence sweep and deployment rule

At M1/T1, behavior-policy concentration is $\tau = 0.35$ and previous-action persistence $\kappa \in \{0.10, 0.25, 0.50, 0.75\}$. BC, CQL, IQL, and Decision Transformer use their confirmed configurations on three new logged-data replicates per level (16,384 training and 2,048 selection episodes). Each selected checkpoint is evaluated on the same 8,192 environment seeds under argmax and sampled deployment. Learner-minus-stochastic-behavior differences pair evaluation seeds within replicate; brackets give two-sided 95% t intervals over three replicates.

Table 23: **S1 mechanism tests at M1/T1.** Shares count decisions; return changes use raw-return units and brackets are 95% intervals when present. Family intervals use Bonferroni correction; paired intervals use logged-data replicates; ranges are descriptive. Reducing CQL alpha to 0.1 yields 42.3% action-0 decisions and raises NS by $+13.7$ $[+8.1, +19.3]$ with a paired t interval (Appendix L); other details appear in Appendices J, J.4, and J.5.

Test and comparator	Method	Action-0 share (%)	Return change [95%]	Interval type
High vs low persistence	CQL	$\geq 99.6\%$ at high	-0.1059 $[-0.1102, -0.1016]$	Family t
	Decision Transformer	$\geq 99.6\%$ at high	-0.1095 $[-0.1293, -0.0897]$	Family t
Reset vs variant constant	CQL	8.86–9.88%	-0.19 to -0.18	Descriptive
	Decision Transformer	8.43–9.00%	-0.19 to -0.18	Descriptive
Sampled vs CPU argmax	CQL	16.11–16.85%	$+0.071$–$+0.079$	Descriptive
	Decision Transformer	44.35–49.07%	$+0.048$–$+0.061$	Descriptive
No previous action vs original	CQL	69.3–73.4%	$+0.0384$ $[+0.0068, +0.0701]$	Paired t
	Decision Transformer	67.2–69.6%	$+0.0729$ $[+0.0600, +0.0857]$	Paired t

Table 24: **Persistence sweep at M1/T1.** Action-0 share counts executed decisions; NS uses the exact M1/T1 constant -0.582707779013 and oracle $+0.043488768905$. The raw-return contrast is learner minus stochastic behavior; behavior-policy rows have no learner contrast.

κ	Deployment	Policy	Action 0	NS	Δ raw return vs. behavior [95%]
0.10	argmax	Behavior policy	12.2%	−12.0	—
0.10	sampled	Behavior policy	12.3%	−13.0	—
0.10	argmax	BC	12.7%	−11.4	+0.0099[−0.0039, +0.0238]
0.10	sampled	BC	12.5%	−12.9	+0.0007[−0.0114, +0.0128]
0.10	argmax	CQL	13.1%	−11.6	+0.0090[−0.0103, +0.0284]
0.10	sampled	CQL	10.7%	−13.7	−0.0040[−0.0321, +0.0242]
0.10	argmax	IQL	14.0%	−9.9	+0.0198[+0.0167, +0.0229]
0.10	sampled	IQL	12.8%	−11.2	+0.0114[−0.0066, +0.0293]
0.10	argmax	Decision Transformer	14.1%	−12.0	+0.0067[−0.0027, +0.0161]
0.10	sampled	Decision Transformer	13.4%	−12.6	+0.0026[−0.0070, +0.0121]
0.25	argmax	Behavior policy	14.8%	−12.3	—
0.25	sampled	Behavior policy	16.2%	−14.7	—
0.25	argmax	BC	17.3%	−13.3	+0.0093[−0.0282, +0.0467]
0.25	sampled	BC	16.5%	−14.7	+0.0000[−0.0406, +0.0406]
0.25	argmax	CQL	18.3%	−13.1	+0.0105[−0.0180, +0.0391]
0.25	sampled	CQL	11.8%	−14.6	+0.0009[−0.0199, +0.0216]
0.25	argmax	IQL	18.0%	−12.3	+0.0154[−0.0181, +0.0489]
0.25	sampled	IQL	15.3%	−13.0	+0.0112[−0.0083, +0.0306]
0.25	argmax	Decision Transformer	15.8%	−10.8	+0.0249[−0.0057, +0.0555]
0.25	sampled	Decision Transformer	16.7%	−13.4	+0.0086[−0.0166, +0.0337]
0.50	argmax	Behavior policy	100.0%	−29.6	—
0.50	sampled	Behavior policy	25.7%	−17.6	—
0.50	argmax	BC	83.1%	−27.6	−0.0622[−0.0928, −0.0315]
0.50	sampled	BC	25.5%	−17.4	+0.0013[−0.0006, +0.0033]
0.50	argmax	CQL	75.5%	−25.0	−0.0462[−0.0943, +0.0018]
0.50	sampled	CQL	13.2%	−15.7	+0.0120[−0.0117, +0.0357]
0.50	argmax	IQL	81.9%	−23.1	−0.0341[−0.0562, −0.0121]
0.50	sampled	IQL	24.3%	−16.4	+0.0075[−0.0038, +0.0189]
0.50	argmax	Decision Transformer	83.4%	−27.9	−0.0644[−0.0980, −0.0308]
0.50	sampled	Decision Transformer	26.0%	−16.5	+0.0069[−0.0102, +0.0241]
0.75	argmax	Behavior policy	100.0%	−29.6	—
0.75	sampled	Behavior policy	45.6%	−21.3	—
0.75	argmax	BC	99.9%	−29.5	−0.0514[−0.0742, −0.0286]
0.75	sampled	BC	47.4%	−21.5	−0.0014[−0.0097, +0.0068]
0.75	argmax	CQL	99.5%	−29.3	−0.0503[−0.0739, −0.0267]
0.75	sampled	CQL	16.5%	−17.2	+0.0256[+0.0119, +0.0393]
0.75	argmax	IQL	99.6%	−29.1	−0.0489[−0.0690, −0.0289]
0.75	sampled	IQL	48.4%	−20.7	+0.0039[−0.0197, +0.0274]
0.75	argmax	Decision Transformer	99.9%	−29.5	−0.0514[−0.0755, −0.0272]
0.75	sampled	Decision Transformer	46.6%	−21.0	+0.0022[−0.0080, +0.0123]

J.3 PRESPECIFIED FACTORIAL CONTRASTS

Writing $q_\tau(a \mid o, p)$ for the concentrated base-action distribution given observation o and previous action p, the implemented behavior policy is

$$\pi_{\tau,\kappa}(a \mid o, p) = 0.98\{(1-\kappa)q_\tau(a \mid o, p) + \kappa \mathbf{1}[a = p]\} + 0.02/12.$$

The 2×2 M1/T1 behavior-policy design sets low/high concentration temperature $\tau = 1.0/0.35$ and low/high persistence $\kappa = 0.10/0.75$, with a 0.02 uniform mixture and previous action reset to 0. The response, reward, action inventory, objectives, and offline checkpoint-selection rules remain fixed. The original S0/S1 laws match low concentration with low persistence/high concentration with high persistence, but none of the original 144 confirmation estimates supplies a factorial cell. Three new role-disjoint data blocks, one prespecified optimization seed index per block, and 8,192 evaluation episodes per fit yield 36 complete fits. Block seeds match across corners within method. There is

no separate optimization-seed variation analysis. The concentration effect averages high-minus-low concentration over persistence. The persistence effect averages high-minus-low persistence over concentration. The interaction is the high-minus-low concentration contrast at high persistence minus the same contrast at low persistence. Each effect is calculated within data block before averaging. Nominal intervals use $t_{0.975,2}$; simultaneous family intervals use the two-sided Bonferroni critical value for all nine total-return contrasts. Table 25 reports all nine contrasts. Intervals spanning zero are inconclusive, not proof of zero effect or additivity.

Across the two concentration levels, persistence-minus-low-persistence clinical and treatment component point contrasts are CQL -0.234946 and $+0.129016$, and Decision Transformer -0.241972 and $+0.132470$. Their total-return contrasts are -0.105929 and -0.109502. Component contrasts have no secondary significance claim. At high persistence, CQL and Decision Transformer have three-block mean action-0 decision frequencies of at least 99.6% at both base temperatures; at low persistence they do not. BC has a smaller descriptive action shift and a family interval for total return that spans zero. Changing persistence also changes conditional behavior-policy support. This factorial alone does not identify copying or mediation; the reset and deployment controls give separate evidence for copying (Appendix J.5).

J.4 Previous-Action Input Ablation

Does previous-action input matter? We removed previous-action input for CQL and Decision Transformer across three M1/T1/S1 replicates. Action-0 decisions fall from above 99% to 67.2–73.4%, a decrease of about 30 percentage points; paired mean return rises $+0.0384$ $[+0.0068, +0.0701]$ and $+0.0729$ $[+0.0600, +0.0857]$, respectively (95% replicate t intervals; Table 23; Appendix J.4). Previous-action input explains part of the repetition; the remaining action-0 share of about 70% shows another pathway remains.

The M1/T1/S1 ablation masks the explicit 12-way previous-action one-hot in all training, selection, and evaluation history positions for CQL and Decision Transformer. It retains the same three confirmation logged datasets, matched payload identities, action labels, reward, normalization, split, checkpoint selection, and evaluation seeds. Each selected fit has 8,192 direct evaluation episodes. Decision Transformer has no separate action tokens; its historical labels were saved before masking and used only as supervision. Other public observations, its reward-fed return-to-go, and dynamics may still reveal previous actions.

Across the three matched replicates, action-0 frequency fell from above 99% to 69.3–73.4% for CQL and 67.2–69.6% for Decision Transformer. Paired ablated-minus-original frequency changes were -0.2798 $[-0.3392, -0.2205]$ (Holm $p = 0.00242$) and -0.3179 $[-0.3496, -0.2862]$ (Holm $p = 0.001074$). Mean return changes were $+0.0384$ $[+0.0068, +0.0701]$ (Holm $p = 0.03471$) and $+0.0729$ $[+0.0600, +0.0857]$ (Holm $p = 0.003349$). The nominal 95% t intervals use three paired logged-data replicates, not episodes, as inference units. These return increases correspond to about $+6.1$ and $+11.6$ M1/T1 NS points; the ablated Decision Transformer mean near -0.695 return is close to the S1 behavior policy at -0.713.

The explicit input affects behavior, yet action-0 frequency remains 67.2–73.4% after masking. This does not isolate a unique copying mechanism or test a different initial previous-action reset. Source records are listed in Appendix N.

J.5 S1 Reset and Deployment Controls

Action 0 or the starting action? We randomized the initial previous action for CQL and Decision Transformer across three M1/T1/S1 replicates. With a uniform initial previous action, CQL and Decision Transformer choose action 0 in 8.4–9.9% of decisions and match the new starting action on over 99% of first decisions; return stays 0.18–0.19 below the variant constant (Table 23; Appendix J.5). The repetition follows the starting setting.

The reset and deployment controls followed a prespecified analysis plan (in the supplement). At M1/T1/S1, each method has three confirmation-matched logged-data replicates and 8,192 direct-evaluation episodes per fit. The reset control draws the initial previous action uniformly for both logging and evaluation; this changes the lifted initial clinical state. Exact variant references remain the best constant -0.582708 and hidden-state oracle $+0.043489$ under core-state projection. Returns

Table 25: **Nine prespecified M1/T1 total-return contrasts.** Mean and nominal and Bonferroni family 95% intervals pair three new data blocks ($df = 2$). Higher return is better; the family interval covers all nine rows.

Method	Effect	Mean	Nominal 95%	Family 95%
BC	Concentration	$+0.0108$	$[-0.0049, +0.0264]$	$[-0.0378, +0.0594]$
BC	Persistence	-0.0527	$[-0.0778, -0.0275]$	$[-0.1307, +0.0254]$
BC	Interaction	-0.0228	$[-0.0550, +0.0094]$	$[-0.1228, +0.0771]$
CQL	Concentration	$+0.0005$	$[-0.0152, +0.0163]$	$[-0.0485, +0.0495]$
CQL	Persistence	-0.1059	$[-0.1073, -0.1045]$	$[-0.1102, -0.1016]$
CQL	Interaction	-0.0002	$[-0.0292, +0.0289]$	$[-0.0903, +0.0899]$
Decision Transformer	Concentration	$+0.0050$	$[-0.0001, +0.0100]$	$[-0.0106, +0.0206]$
Decision Transformer	Persistence	-0.1095	$[-0.1159, -0.1031]$	$[-0.1293, -0.0897]$
Decision Transformer	Interaction	-0.0100	$[-0.0181, -0.0018]$	$[-0.0352, +0.0153]$

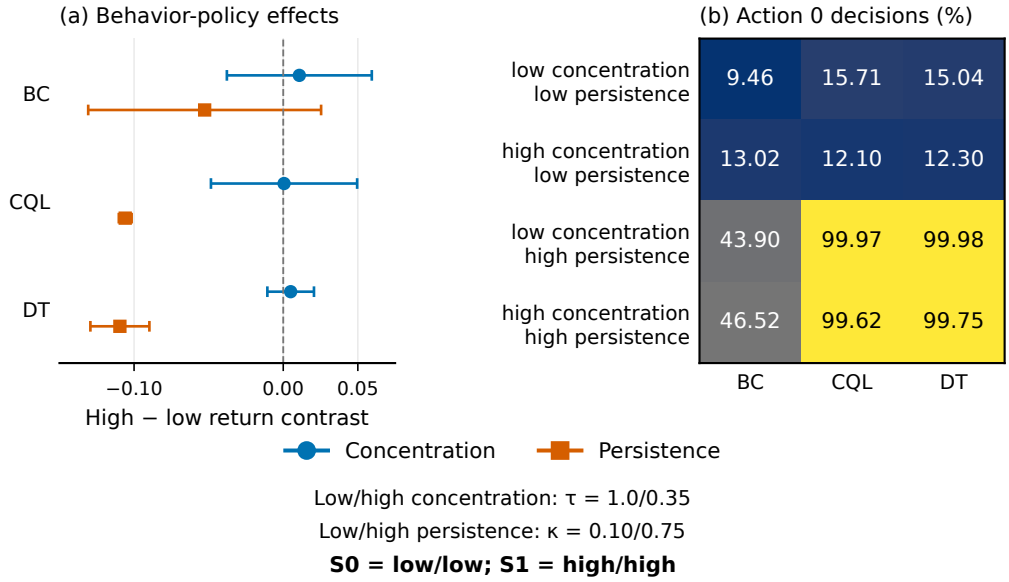

Figure 7: **Behavior-policy concentration and persistence at M1/T1.** Low/high concentration uses $\tau = 1.0/0.35$; low/high previous-action persistence uses $\kappa = 0.10/0.75$. Left: return contrasts with Bonferroni intervals across paired data blocks. Right: descriptive action-0 frequencies. Table 25 gives exact return contrasts.

remain below the variant constant. The deployment control uses original-S1 CPU refits: Decision Transformer samples its legal-action distribution; CQL samples temperature-1 Boltzmann probabilities over per-decision normalized legal Q-values. Each sampled policy is paired with its own CPU argmax evaluation; selected CPU checkpoints differ from GPU confirmation checkpoints.

With a random starting action, action-0 share falls to 8.4–9.9%, and the first action matches the starting action in over 99% of episodes (chance $1/12$). Repetition follows the starting action rather than action 0 specifically. Sampling improves return but does not establish a unique internal copying pathway. The prespecified analysis plan covered CQL and Decision Transformer, matching the earlier previous-action ablation; IQL and BC were outside these controls.

K Evaluation-Time Action and Behavior-Policy Support Audit

A separate analysis replayed all 144 selected Respiratory confirmation policies over their original episode sets of 8,192 evaluations each. Every episode's return and reward components agreed with

Table 26: **S1 reset and deployment controls by logged-data replicate.** Reset return is mean direct return in the new variant; anchor agreement counts first decisions. Original-S1 deployment returns compare matched CPU refits. Shares count executed decisions.

		Random reset variant			Original S1 deployment			
Method	Rep.	Action 0	Anchor	Return	Argmax 0	Sample 0	Argmax V	Sample V
CQL	1	8.86%	99.58%	-0.7652	99.23%	16.11%	-0.7652	-0.6941
CQL	2	9.88%	99.72%	-0.7733	99.47%	16.48%	-0.7682	-0.6903
CQL	3	8.94%	99.73%	-0.7684	99.82%	16.85%	-0.7658	-0.6871
DT	1	8.43%	99.89%	-0.7675	99.92%	44.35%	-0.7663	-0.7184
DT	2	9.00%	99.68%	-0.7746	99.95%	49.07%	-0.7686	-0.7165
DT	3	9.00%	99.55%	-0.7689	99.80%	46.36%	-0.7676	-0.7068

the retained evaluation within 10^{-10}, and decision counts matched exactly. No learner was retrained or retuned. The retained aggregate action audit covers 144 fits, 48 cells, and 24 paired S contrasts; source records appear in Appendix N. Only fit-level aggregate action statistics were saved.

Action frequency weights executed decisions equally. *Minimum pair* denotes dense action 0, the lowest supported PEEP/FiO_2 pair. Its normalized action intensity is zero, but this label does not indicate an absence of ventilation or oxygen. *Low intensity* includes actions at or below the frozen threshold 0.4481614068, the 25th percentile of strictly positive Respiratory action intensities. Reported intervals for action metrics use the three independent logged-data replicates and do not estimate optimization-seed variation or adjust for multiplicity.

For each method, M/T cell, and replicate, the paired S1-minus-S0 calculation compares the same evaluated method and then averages the four M/T contrasts within each replicate. CQL, IQL, and Decision Transformer increase minimum-pair frequency by 86.0, 86.5, and 84.7 percentage points. Their corresponding nominal 95% replicate intervals are $[80.7, 91.4]$, $[83.8, 89.2]$, and $[82.3, 87.2]$ points. Their treatment-component changes are $+0.131$, $+0.126$, and $+0.132$; combined simulated event-component changes are -0.229, -0.237, and -0.235. The total-return changes are -0.098, -0.111, and -0.103. The supplement provides 144 per-fit action summaries in `analysis/v1_calibration/respiratory_eval_action_trace_v1/per_fit_action_summary.csv`; the 24 method-by-M-by-T S1-minus-S0 action contrasts can be recomputed from these summaries. Replicate intervals and 12-way action-distribution total-variation distances require the corresponding aggregation rather than a separate supplementary contrast table.

At M1/T1/S1, CQL, IQL, and Decision Transformer use action 0 in 99.5%, 99.6%, and 99.9% of decisions. Full-record POMCP in a separate run uses it in 0.07% of decisions. For each learner-selected action, the audit also scores both frozen behavior-policy laws on the public history generated by that learner. At M1/T1/S1, the mean probability assigned by the S1 behavior policy to these three methods' chosen actions is approximately 0.766–0.767. Below-0.01 selected-propensity fractions are 0 for CQL, 2.4474×10^{-5} for IQL, and 8.2946×10^{-6} for Decision Transformer. The fifth-percentile propensity is approximately 0.73667 for each, matching the S1 repeat-action contribution $0.98(0.75) + 0.02/12$. The environment initializes previous action to 0. Repeating that action therefore contributes strongly to selected-action propensity without establishing an independent preference for action 0. This action audit alone cannot distinguish copying from preference; the prespecified reset and deployment controls support copying for CQL and Decision Transformer (Appendix J.5). IQL has no corresponding reset control. Those local propensities are not whole-trajectory importance weights. The logged training data themselves change between S0 and S1: across M/T cells, minimum-pair frequency increases by 31.7 points and mean normalized intensity decreases by 0.191. Learner-action, behavior-policy, and outcome shifts are associated within the specified simulator; they do not isolate a single mediator or a clinical effect.

L Optimistic Tuning Sensitivity

Does weaker CQL conservatism matter? We compared tuned CQL alpha 0.1 with the default alpha across three M1/T1/S1 replicates on a separate episode set. Tuning reduces action-0 decisions to 42.3% and gains $+13.7$ NS $[+8.1, +19.3]$ (Table 23; Appendix L). For CQL and Decision

Table 27: **Respiratory evaluation-time action behavior and reward differences.** Eight M/T/S cells are equally weighted within each of three independent data-replicate blocks. Minimum and low-intensity columns weight decisions; reward differences subtract the exact best constant. Values are descriptive point estimates; replicate intervals and all 12-action distributions are in the source aggregate tables.

Method	Intensity	Minimum %	Low %	Event Δ	Treat. Δ	Return Δ
BC	0.451	25.6	51.9	-0.125	$+0.012$	-0.113
DQN/FQI	0.626	2.4	23.3	-0.016	-0.034	-0.050
CQL	0.265	56.6	73.2	-0.191	$+0.055$	-0.136
IQL	0.264	56.4	71.7	-0.187	$+0.058$	-0.129
Decision Transformer	0.264	57.5	72.2	-0.189	$+0.055$	-0.134
MedDreamer	0.664	6.4	15.4	$+0.036$	-0.052	-0.016

Transformer where tested, the controls support repetition of the starting setting under persistent logs, amplified by argmax. Previous-action input and, for CQL, conservatism strength each explain part of the lock-in.

The tuning sensitivity followed a prespecified analysis plan (in the supplement). Three new logged-data replicates per method and S condition each used 16,384 training and 2,048 offline-selection episodes. CQL alpha was selected from $\{0.1, 1, 3, 10\}$; DQN/FQI learning rate from $\{0.00009, 0.0003, 0.0009\}$. A separate simulator episode set of 2,048 episodes selected each configuration, then a disjoint reporting episode set of 8,192 episodes evaluated it. Such simulator-evaluation selection is unavailable in offline practice. The gains are optimistic grid-based sensitivity estimates, not deployable gains or formal statistical upper bounds. NS uses nominal M1/T1 exact anchors; paired differences use nominal 95% t intervals across three logged-data replicates, without multiplicity adjustment.

In S1, tuned CQL chose action 0 in 42.34% of decisions and tuned DQN/FQI in 1.05%. All tuned scores remain below the shared-seed POMCP references of 31.5–35.3 NS. These tested hyperparameters do not explain the learner–planner gap; the control does not assess other settings or selection without simulator evaluation.

M Source-Cohort Action Persistence

The descriptive analysis uses only the MIMIC-IV v3.1 Respiratory source training cohort and the existing dense PEEP/FiO_2 action codec. Each half-open four-hour bin requires direct observations of both channels. An eligible pair has consecutive transition indices within one ICU episode and passes the original target-membership and temporal-repair rules; an observed pair additionally has a mapped dense action in both bins. Repeat means equality of the two dense actions, without filling or bridging missing bins. Of 155,589 eligible potential pairs, 74,503 were observed (47.88% coverage) across 12,110 subjects and 14,043 episodes; 60,551 observed pairs repeated (81.27%, 95% interval 80.94–81.62%). The interval uses 2,000 subject-cluster bootstrap replicates and weights observed pairs equally. For the specified synthetic behavior policies, 2,048 development-seed episodes per condition give 3,138/13,358 S0 and 10,154/12,947 S1 repeats at M1/T0 four-hour transitions (23.49/78.43%), and 1,808/8,008 S0 and 6,395/8,139 S1 repeats at M1/T1 variable-interval decisions (22.58/78.57%). Recorded repeats may include unchanged-setting documentation; the observed pairs cannot establish continuous persistence between observations. The comparison is source-cohort context, not behavior-policy calibration, evidence that S1 reproduces practice, or clinical validation.

N Source Identity and Evidence Roles

The supplementary release is `release/respiratory_v1/`; its `REPRODUCE.md` gives commands and the scope of author reruns. Its `B1_REPORT.md` and `FRESH_ENV_RECEIPT.json` retain author fresh-environment qualification: exact references and scoped full CPU CQL and DQN/FQI fits matched recorded results within stated tolerances, while selected CPU tensors differed from recorded CUDA tensors. The Respiratory aggregate report and `analysis/v1_calibration/`

Table 28: **Optimistic M1/T1 tuning sensitivity.** Delta compares tuned with frozen configuration on paired reporting episodes; S1 action-0 share, reported in the following paragraph, is decision-weighted. Shared-seed POMCP references score 31.5–35.3 NS.

S	Method	Selected	Tuned NS	Frozen NS	Δ NS [95%]
S0	CQL	$\alpha = 0.1$	-2.90	-10.71	$+7.81\ [-1.77, +17.40]$
S0	DQN/FQI	0.0003	7.36	7.36	$0.00\ [0.00, 0.00]$
S1	CQL	$\alpha = 0.1$	-15.86	-29.52	$+13.67\ [+8.05, +19.28]$
S1	DQN/FQI	0.00009	-6.11	-12.29	$+6.18\ [-10.01, +22.37]$

`respiratory_final_v1/artifacts/source_manifest.json` are in the supplement; the manifest binds reported tables to exact files. The confirmed learner settings are in `configs/offline_suite/method_manifest_v1_final_confirmation.json`. The numerical simulator laws are under `src/ehrgym/_native/` and `src/ehrgym/resources/`. The corrected aggregate-input SHA-256 is `52787b58a5f9cec56d29efd7384fd09c92bbb5f53e49c56f9cba1ac1d1a5ec09`. The manuscript's numerical specification uses the same synthetic transition, reward, action, horizon, and M/T/S laws as the release; the corrected empirical input mapping affects the observation background. Computational consistency checks do not establish clinical validity.

O Closest-Work Comparison

The following comparison identifies where EHRGym's combination of inputs, references, and diagnostics fits among relevant clinical and offline-RL resources.

Table 29: **Source-bound positioning of the Respiratory study.** The cited works' tasks and reported evaluation roles are summarized at the level needed for this comparison.

Work	Reported setting and relation to the present comparison
Sepsis policy studies (Raghu et al., 2017; Komorowski et al., 2018)	Learn treatment policies from retrospective ICU sepsis records. Their outcomes motivate care in policy interpretation; EHRGym evaluates a non-sepsis synthetic Respiratory law with exact and executed diagnostic references.
ICU-Sepsis (Choudhary et al., 2024)	A tabular sepsis benchmark MDP built from real medical data. EHRGym instead uses a partially observed Respiratory action and acquisition interface and tests full-record planner recoverability before learner diagnosis.
EpiCare (Hargrave et al., 2024)	Medical POMDPs across eight instances with a nonprivileged symptom-history Standard-of-Care reference and greedy privileged-state reference. It tests offline learners and OPE, training-data sources, and data-size restrictions. EHRGym's Respiratory case combines aggregate EHR acquisition inputs, exact finite-horizon references, an executed full-record planner, and M/T/S action audit.
DTR-Bench (Luo et al., 2024)	Four in silico treatment settings, including sepsis, with noise, partial-observation and missing-data tests. The present Respiratory analysis binds EHR aggregate recording inputs to a controlled M/T/S and reference-ladder protocol.
B4MRL (Linial et al., 2024)	Generic offline-data and simulator discrepancies, including partial observability and action/state mismatch. EHRGym tests those kinds of algorithmic limitations through a specified EHR-motivated decision task and executed action audit.
MedGym (Wang et al., 2026)	Continuous-time patient evolution and irregular measurement/intervention timing in a configurable clinical-data-derived treatment benchmark. Its timing design is relevant prior work; EHRGym's T1 is a discrete hidden-burden-dependent hold.
Copycat actions (Wen et al., 2020)	Observation-history imitation can exploit correlation with a previous action. EHRGym's S1 controls support repetition copying for CQL and Decision Transformer in this synthetic cell, without establishing their internal representation or a general mechanism.

Table 29: **Source-bound positioning of the Respiratory study (continued).**

Work	Reported setting and relation to the present comparison
MedDreamer (Xu et al., 2025)	An EHR-oriented model-based method evaluated on sepsis and ventilation data. The MedDreamer-style entrant here is a compact benchmark adaptation evaluated in a synthetic task; it is not an official reproduction or a clinical outcome estimate.