WHO SHOULD LEARN MORE, AND WHEN? CONFLICT-GATED CREDIT ROUTING FOR OFFLINE RL WITH COMPETING VALUE BRANCHES

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical reinforcement learning (RL) has largely optimized treatment for a single disease, leaving cross-disease coordination in critically ill patients underexplored. Sepsis–acute kidney injury (AKI) exemplifies this challenge, where the same intervention may benefit one disease while worsening the other, and the clinically dominant disease may shift over time. This requires a policy to decide, at each moment, whose priority weight (i.e., Sepsis's or AKI's) should dominate learning, and how their competing treatment preferences are reconciled into a single action. We provide the first offline multi-agent reinforcement learning (MARL) formulation of comorbid care and introduce our model BIFOCAL, which coordinates disease-conditioned value branches through **dynamic credit routing**: a *telescope* (future deterioration risk) sets which disease deserves priority, a *magnifying glass* (present treatment conflict) determines when that priority takes effect, and their coupling decides which disease dominates learning and reconciles preferences into a single action. Across 4 diverse ICU datasets from U.S. and China comprising **87,566** trajectories with **1,285,977** decision points, BIFOCAL consistently achieves higher estimated policy value than clinician behavior. On MIMIC-IV, it further outperforms baselines in the comorbid cohort and two disease-specific methods in disease-only cohorts under 5 off-policy and 2 action-consistency metrics. Its lead on single-disease cohorts suggests that BIFOCAL's utility extends beyond comorbidity to single-disease care. Furthermore, a **blinded physician evaluation** of 30 comorbid cases over six dimensions supports its clinical potential. These results establish BIFOCAL as the first large-scale offline MARL framework for cross-disease treatment coordination in critical care.

1 INTRODUCTION

Sepsis is one of the most common and lethal causes of critical illness in intensive care, accounting for an estimated 11 million deaths annually worldwide Rudd et al. (2020); Fleischmann-Struzek et al. (2020). Acute kidney injury (AKI) is a frequent complication of sepsis, occurring in approximately 40–50% of septic patients, and *coexisting sepsis and AKI* (Sepsis-AKI), carries substantially higher mortality than isolated sepsis or isolated AKI Takeuchi et al. (2025). Beyond its prevalence, Sepsis-AKI poses a particularly challenging treatment problem because interventions intended to improve one physiological objective may simultaneously worsen another. For example, fluid resuscitation may restore perfusion in sepsis while worsening fluid burden in AKI Payen et al. (2008). As a result, effective management requires continuous adaptation to the patient's evolving state rather than a fixed therapeutic protocol.

Offline reinforcement learning (RL) is well suited to such sequential decision problems because it learns treatment policies from fixed retrospective ICU trajectories without online interaction while optimizing long-term clinical consequences rather than only immediate physiological changes (Levine et al., 2020). However, most existing clinical RL studies formulate treatment as a single-disease decision problem with one undifferentiated value function (Komorowski et al., 2018; Zhang et al., 2022; Peine et al., 2021; Xu et al., 2026; Wu et al., 2023). Such formulations do not explicitly represent the distinct treatment perspectives of coexisting diseases or determine which disease-specific branch should receive greater influence at each patient state. Importantly, the clini-

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.

cally dominant disease and the corresponding treatment priority can change over time. For example, treatment may need to initially prioritize stabilizing sepsis when circulatory failure is the greatest concern, but later shift toward protecting kidney function as AKI worsens. If the patient's circulatory status deteriorates again, treatment priority may return to sepsis.

A natural extension is multi-objective RL, which can encode both sepsis and AKI objectives within a single optimization framework Abels et al. (2019); Yang et al. (2019). Yet these approaches still rely on a single decision maker that balances objectives internally. They do not explicitly represent the distinct clinical perspectives of sepsis-oriented and AKI-oriented treatment processes, nor do they provide a mechanism for determining which disease should exert greater influence at a particular patient state. Existing offline multi-agent reinforcement learning (MARL) methods improve coordination among agents via value decomposition or credit assignment Yang et al. (2021); Shao et al. (2023); Wang et al. (2023; 2025); Zhou et al. (2025), but their credit mechanisms quantify each agent's contribution to an already aggregated objective. In Sepsis-AKI, however, the key question is upstream of aggregation: before the final decision, the model must decide which disease deserves higher priority given the patient's state.

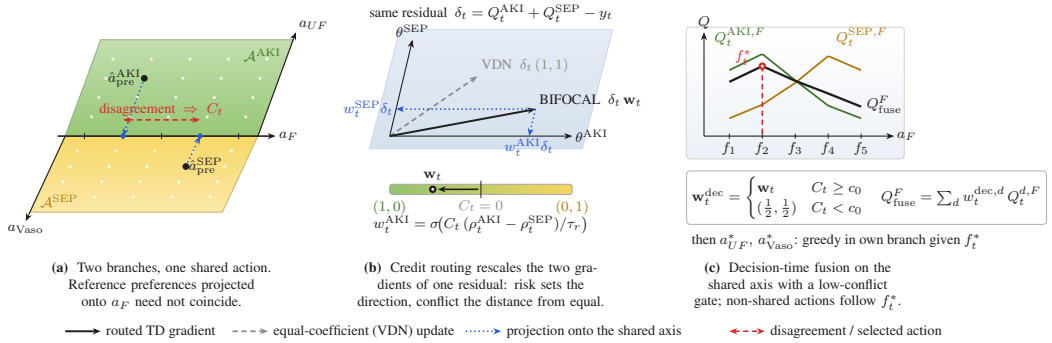
This observation leads to our central insight: disease priority should emerge from **two complementary signals** rather than from a fixed weighting scheme. First, future deterioration risk (acting as a telescope) sets the direction of priority by identifying which disease is more likely to worsen in the near future. Second, present treatment conflict (acting as a magnifying glass) controls the strength of that priority, because asymmetric coordination is needed only when disease-specific treatment preferences diverge. For example, when both agents favor additional fluid, risk-based reweighting is unnecessary; when septic hypoperfusion coexists with AKI-related fluid overload, the agents may disagree sharply, and future risk should determine the direction of reweighting. Thus, risk identifies which disease should be favored, conflict determines whether and how strongly it should be favored, and together they allocate influence in learning and action selection.

Based on this principle, we study Sepsis–AKI, a prevalent and archetypal comorbidity with frequently conflicting treatment goals, and formulate its coordinated treatment as an offline MARL problem. To our knowledge, this is the first large-scale, cross-dataset investigation of cross-disease coordination in critical care. We introduce BIFOCAL, named for its dual focus on these complementary signals, with four key contributions:

- **Offline MARL formulation for Sepsis–AKI coordination.** We model sepsis-oriented and AKI-oriented treatment processes as separate agents that learn distinct value functions from shared ICU trajectories while producing a single executable treatment.
- **Dynamic risk–conflict credit routing.** We introduce a two-signal mechanism in which future deterioration risk sets the direction of disease priority, present treatment conflict controls its activation, and their interaction allocates TD updates during learning and adaptively fuses Q-values during action selection.
- **Comprehensive evaluation.** On MIMIC-IV, BIFOCAL outperforms five general-purpose RL baselines on all five OPE metrics in the comorbid cohort and two disease-specific baselines on most OPE metrics in disease-only cohorts. It also achieves higher estimated policy value than clinician behavior across **4** U.S. and Chinese ICU datasets comprising **87,566 trajectories** and **1,285,977** decision points, supporting its consistency across diverse clinical settings.
- **Clinical utility assessment.** Six ICU physicians evaluated BIFOCAL on **30** real comorbid cases across six dimensions in a blinded evaluation. Their ratings support the clinical plausibility of BIFOCAL's learned disease priorities and coordinated treatments.

2 METHOD

We consider offline value learning with two disease-conditioned branches that share one action dimension and one terminal return. Throughout, $d \in \{\mathrm{AKI}, \mathrm{SEP}\}$ indexes the branch and t the decision step. Sec. 2.1 fixes the value representation and shows that additive decomposition leaves one degree of freedom unspecified; Sec. 2.2 closes it with a factorized routing rule and characterizes what that rule does and does not change about the learning problem; Sec. 2.3 carries the same two



(a) Two branches, one shared action. Reference preferences projected onto a_F need not coincide.

(b) Credit routing rescales the two gradients of one residual: risk sets the direction, conflict the distance from equal.

(c) Decision-time fusion on the shared axis with a low-conflict gate; non-shared actions follow f_t^*.

Figure 1: **The shared-action coordination problem and BIFOCAL's response. (a)** The AKI and Sepsis branches own disjoint sub-actions (a_{UF}, a_{Vaso}) but both act on the shared fluid axis a_{F}; their frozen reference preferences projected onto that axis need not coincide, and the size of that gap defines the conflict signal C_t. **(b)** Both branches receive one additive residual δ_t. Standard VDN applies it with equal coefficients; BIFOCAL rescales the two gradients by $\mathbf{w}_t$, where predicted risk sets the direction of the split and C_t sets how far it departs from $(1/2, 1/2)$, recovering VDN exactly at $C_t = 0$. **(c)** At decision time the same two signals fuse the branch Q-profiles along the shared axis under a low-conflict gate; the non-shared actions are then chosen greedily within each branch given $a^*_{t,\text{F}}$. The complete architecture, including cohort construction, risk-module pretraining, router supervision, and anchoring, is given in Appendix Fig. 12.

signals to action selection and quantifies the resulting operator mismatch. Fig. 1 summarizes the problem structure and the two mechanisms it motivates; the full pipeline diagram is deferred to Appendix Fig. 12.

2.1 Disease-Conditioned Value Branches under a Shared Action

Comorbid treatment involves competing disease considerations because AKI and sepsis can favor different uses of a shared intervention. For example, fluid may support septic circulation while increasing AKI-related volume burden. A single value function obscures these disease-oriented perspectives, so we use one value branch for each disease. Let $d \in \{\text{AKI}, \text{SEP}\}$ denote the disease identity, with SEP denoting sepsis, and t the decision step. At step t, agent d receives its disease-specific state s_t^d and maintains a trainable action-value function $Q^d(s_t^d, a_t^d)$ over its disease-specific sub-action a_t^d.

Let $\mathcal{A}_{\text{F}}$, $\mathcal{A}_{\text{UF}}$, and $\mathcal{A}_{\text{Vaso}}$ denote the discrete fluid, ultrafiltration, and vasopressor action sets. The disease-specific spaces are $\mathcal{A}^{\text{AKI}} = \mathcal{A}_{\text{F}} \times \mathcal{A}_{\text{UF}}$ and $\mathcal{A}^{\text{SEP}} = \mathcal{A}_{\text{F}} \times \mathcal{A}_{\text{Vaso}}$, yielding $\mathcal{A}_{\text{joint}} = \mathcal{A}_{\text{F}} \times \mathcal{A}_{\text{UF}} \times \mathcal{A}_{\text{Vaso}}$. At step t, the joint state is $\mathbf{s}_t = (s_t^{\text{AKI}}, s_t^{\text{SEP}})$, and the treatment $\mathbf{a}_t = (a_{t,\text{F}}, a_{t,\text{UF}}, a_{t,\text{Vaso}})$ induces $a_t^{\text{AKI}} = (a_{t,\text{F}}, a_{t,\text{UF}})$ and $a_t^{\text{SEP}} = (a_{t,\text{F}}, a_{t,\text{Vaso}})$. The agents share a reward r_t, which is zero for nonterminal transitions and equals the binary 90-day survival outcome at termination, indicated by $\text{done}_t \in \{0, 1\}$. Given the offline dataset $\mathcal{D}_{\text{com}} = \{(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}, \text{done}_t)\}$, we formulate comorbid treatment as a cooperative multi-agent Markov decision process and optimize a joint action-value function Q through TD learning. Its target is $y_t = r_t + \gamma(1 - \text{done}_t) \max_{\mathbf{a}' \in \mathcal{A}_{\text{joint}}} Q(\mathbf{s}_{t+1}, \mathbf{a}')$, where $\gamma \in [0, 1]$ is the discount factor, and the objective is $\mathcal{L}_{\text{TD}} = \mathbb{E}_{\mathcal{D}_{\text{com}}}[(Q(\mathbf{s}_t, \mathbf{a}_t) - y_t)^2]$. Thus, the branches remain disease-specific, while the common comorbid outcome produces their shared TD residual.

To initialize both branches with disease-specific expertise, we first pretrain a Q-learning agent for each disease $d \in \{\text{AKI}, \text{SEP}\}$ on single-disease trajectories, obtaining the value function Q^d_{pre} and encoder representation $h^d_{t,\text{pre}}$. Each comorbid value function Q^d is initialized from Q^d_{pre}, while the pretrained model is retained as a fixed reference to measure disease disagreement and limit drift from disease-specific expertise during joint training. The two agents are then jointly optimized for the comorbid outcome using value-decomposition networks (VDNs) (Sunehag et al., 2018), where $Q_{\text{VDN}}(\mathbf{s}_t, \mathbf{a}_t) = Q^{\text{AKI}}(s_t^{\text{AKI}}, a_t^{\text{AKI}}) + Q^{\text{SEP}}(s_t^{\text{SEP}}, a_t^{\text{SEP}})$ instantiates the TD objective above. Unlike QMIX's nonlinear state-conditioned mixer, this additive form preserves explicit disease-conditioned value branches. However, standard VDN applies the same joint TD residual to

both without determining which branch should learn more from it. We make this degree of freedom explicit before closing it.

2.2 Dynamic Risk–Conflict Credit Routing

Additive decomposition determines the forward value but not how the shared residual is distributed: any per-branch coefficient vector on the simplex leaves Q_{VDN} unchanged while altering the relative rate at which the branches move. Standard VDN fixes that vector to $(1/2, 1/2)$ in every state, which is a choice rather than a consequence of the objective. We instead make it state-dependent, and specify it with two deliberately separate factors: a *direction* that orders the branches, and an *activation* that scales how far the coefficients may leave the equal-credit point. Keeping them separate is the design hypothesis, not a convenience — an ordering applied where the branches already agree perturbs a consensus without arbitrating anything, while a disagreement signal without an ordering reports that arbitration is needed but not in whose favor — and it is what the Uniform, Risk-only, and Shuffled-risk ablations of Sec. 3.4 are designed to test.

We govern these decisions with two complementary signals. The first, future deterioration risk, acts as a ***telescope***: it looks beyond the current state and estimates which disease trajectory is more likely to dominate going forward, so that a transition is attributed more strongly to the agent whose condition is gaining momentum. Yet risk alone indicates potential disease dominance, not whether the agents actually disagree over the current treatment. The second signal, present-state conflict, acts as a ***magnifying glass***: when the two agents' preferred treatments diverge, arbitration is necessary, but when they already agree, forcing risk-based reallocation would perturb a stable consensus. Conflict, however, is non-directional, it reveals *that* the agents diverge but not *which* disease to favor. Only their interaction produces fully specified credit, future risk indicates the direction and present conflict controls its activation.

Direction: Future Risk (a Telescope). Because current severity may not reveal which disease will worsen, we pretrain a future-representation module using the joint state $\mathbf{s}_t$ from Sec. 2.1. For each valid step, the history window $H_t = (\mathbf{s}_{t-3}, \ldots, \mathbf{s}_t)$ covers four 4-hour steps, while the future window $F_t = (\mathbf{s}_{t+1}, \ldots, \mathbf{s}_{t+6})$ covers the subsequent six steps (24 hours; the horizon is selected in Appendix Fig. 9). Risk-model pretraining has two phases. *During representation learning*, a history encoder maps H_t to the predicted future representation $\hat{z}_t^{\mathrm{future}}$, while a future encoder maps F_t to the target representation z_t^{future}. The prediction loss $\mathcal{L}_{\mathrm{pred}}$ aligns these representations, and the reconstruction loss $\mathcal{L}_{\mathrm{recon}}$ preserves sufficient information in z_t^{future} to reconstruct F_t. This enables the history encoder to anticipate future clinical evolution rather than merely summarize the current condition. *During risk supervision*, a **Future-Risk Predictor** applies two sigmoid heads to $\hat{z}_t^{\mathrm{future}}$, producing disease-specific risk scores $\rho_t^{\mathrm{AKI}}, \rho_t^{\mathrm{SEP}} \in [0, 1]$. These scores are supervised by continuous deterioration targets derived from future renal and sepsis-related measurements selected according to KDIGO and Sepsis-3/SSC criteria (Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group, 2012; Singer et al., 2016; Evans et al., 2021). After pretraining, the history encoder and risk heads are frozen and use only H_t during joint-policy training and inference; the future encoder and reconstruction decoder are discarded. The scores determine the disease-priority direction for credit routing, while present conflict controls its strength. Details of the guideline-anchored deterioration targets and the two-stage risk-model training are provided in Appendices B.4–B.5, and the marker sets and cohort statistics in Appendix Table 6.

Activation: Present-Conflict Gating (a Magnifying Glass). Future risk indicates which disease may worsen, but unequal routing is needed only when disease-specific treatment preferences conflict. We therefore introduce a **Conflict Router** to estimate present-state conflict from compact agent outputs. For each disease d, the router input includes the pretrained disease-specific encoder representation $h_{t,\mathrm{pre}}^d$ defined in Sec. 2.1 and the greedy sub-action proposal $\hat{a}_{t,\mathrm{pre}}^d = \arg\max_{a^d} Q_{\mathrm{pre}}^d(s_t^d, a^d)$. The router has parameters η and produces a sigmoid conflict score $C_t = g_\eta(h_{t,\mathrm{pre}}^{\mathrm{AKI}}, h_{t,\mathrm{pre}}^{\mathrm{SEP}}, \hat{a}_{t,\mathrm{pre}}^{\mathrm{AKI}}, \hat{a}_{t,\mathrm{pre}}^{\mathrm{SEP}}) \in [0, 1]$. To supervise the router, we derive a conflict target from the frozen disease-specific value functions Q_{pre}^d in Sec. 2.1. Because fluid is the only action dimension shared by both agents, disagreement over fluid provides the most direct measure of treatment conflict: fluid may support septic perfusion while increasing AKI-related fluid burden. We therefore utilize this to quantify the conflict. For each fluid level $f \in \mathcal{A}_{\mathrm{F}}$, we maximize Q_{pre}^d over the disease-specific nonshared dimension—ultrafiltration for AKI and vasopressor for Sepsis;

and apply a softmax across fluid levels to obtain $p_t^d(f)$. This yields the following notion of conflict, defined entirely in terms of the frozen references and therefore computable before joint training begins.

Definition 1 (Shared-action conflict)**.** *Let p_t^d be branch d's reference preference distribution over the shared action set $\mathcal{A}_{\mathrm{F}}$, obtained by maximizing Q^d_{pre} over its non-shared dimension and applying a softmax across shared-action levels. The shared-action conflict at state $\mathbf{s}_t$ is $\tilde{C}_t = \mathrm{JSD}(p_t^{\mathrm{AKI}}, p_t^{\mathrm{SEP}})/\log 2 \in [0,1]$, with $\tilde{C}_t = 0$ iff the two branches induce identical preferences over $\mathcal{A}_{\mathrm{F}}$.*

We train the router with $\mathcal{L}_{\mathrm{conf}} = (C_t - \tilde{C}_t)^2$ so that C_t approximates Def. 1 from inputs available at decision time. C_t estimates *reference-policy disagreement*, not an independently validated clinical label; the frozen reference also distinguishes it from post-training branch disagreement, though empirically the router's output correlates 0.77 with current-branch disagreement versus 0.52 with the frozen reference (Appendix A.3).

Allocation: The Risk $\times$ Conflict Interaction. We combine the risk vector $\boldsymbol{\rho}_t = (\rho_t^{\mathrm{AKI}}, \rho_t^{\mathrm{SEP}})$ with the predicted conflict through **Risk–Conflict Weighting**. The credit weights are

$$\mathbf{w}_t = (w_t^{\mathrm{AKI}}, w_t^{\mathrm{SEP}}) = \mathrm{softmax}(C_t\, \boldsymbol{\rho}_t/\tau_r)\,, \tag{1}$$

where the softmax is taken over the two diseases and $\tau_r > 0$ controls sensitivity to their risk difference. The weights sum to one: they equal $(1/2, 1/2)$ when $C_t = 0$, remain near equal under low conflict, and increasingly favor the disease with the larger risk score as conflict rises. For two branches Eq. 1 reduces to $w_t^{\mathrm{AKI}} = \sigma\big(C_t(\rho_t^{\mathrm{AKI}} - \rho_t^{\mathrm{SEP}})/\tau_r\big)$, which makes the factorization explicit: the sign of the risk difference sets *which* branch is favored, while C_t scales *how far* the weights may travel from the equal-credit point.

For the observed sub-action, let $Q_t^d = Q^d(s_t^d, a_t^d)$. We define $Q^d_{t,\mathrm{route}} = \mathrm{sg}(Q_t^d) + w_t^d[Q_t^d - \mathrm{sg}(Q_t^d)]$, where $\mathrm{sg}(\cdot)$ is the stop-gradient operator. Therefore, $Q^d_{t,\mathrm{route}}$ equals Q_t^d in the forward pass but scales its TD gradient by w_t^d. Using the target y_t defined in Sec. 2.1, the routed TD loss is $\mathcal{L}_{\mathrm{TD}}^{\mathrm{route}} = (Q^{\mathrm{AKI}}_{t,\mathrm{route}} + Q^{\mathrm{SEP}}_{t,\mathrm{route}} - y_t)^2$, whose branch gradient, with a detached target and frozen weights, is

$$\nabla_{\theta_d}\mathcal{L}_{\mathrm{TD}}^{\mathrm{route}} = 2\,\delta_t\, w_t^d\, \nabla_{\theta_d} Q_t^d, \qquad \delta_t = Q_t^{\mathrm{AKI}} + Q_t^{\mathrm{SEP}} - y_t. \tag{2}$$

Proposition 1 (Routing preserves the additive fixed point)**.** *Let the TD target be detached and the routing weights be frozen, with $w_t^d > 0$ and $\sum_d w_t^d = 1$. Then for every state the routed gradient of Eq. 2 is a strictly positive, state-dependent rescaling of the corresponding VDN gradient, and the two objectives share the same stationarity condition $\delta_t = 0$. Moreover $\mathbf{w}_t \to (1/2, 1/2)$ as $C_t \to 0$, so uniform credit is recovered exactly.*

Remark 1. Prop. 1 delimits the claim in both directions: routing changes the optimization path and each branch's effective learning rate, but not the set of additive fixed points, and it does not estimate a branch's causal contribution to the outcome. "Equal credit" means an equal gradient *coefficient*, not equal gradient norms. Since $C_t = 0$ recovers VDN exactly, the method is a strict generalization of its backbone with the gate controlling how far it may depart — so Risk-only routing, which applies the same ordering even where nothing is contested, is the diagnostic ablation rather than a weaker variant.

The full objective is $\mathcal{L}_{\mathrm{MARL}} = \mathcal{L}_{\mathrm{TD}}^{\mathrm{route}} + \lambda_{\mathrm{anchor}}(1 - C_t)\sum_{d\in\{\mathrm{AKI,SEP}\}} \mathcal{L}^d_{\mathrm{anchor}}$, where $\lambda_{\mathrm{anchor}} \geq 0$ controls anchoring strength and $\mathcal{L}^d_{\mathrm{anchor}} = \mathrm{KL}(\pi^d_{\mathrm{pre}}(\cdot \mid s_t^d) \| \pi^d(\cdot \mid s_t^d))$. Here, π^d_{pre} and π^d are the action distributions induced by the frozen pretrained and current Q-functions, respectively. The anchor is strongest at $C_t = 0$ and vanishes at $C_t = 1$. During MARL training, the risk model and router parameters are frozen, and no gradient is propagated through $\boldsymbol{\rho}_t$, C_t, or $\mathbf{w}_t$. The pseudocode of the complete learning procedure is given in Algorithm 1 (Appendix B.2).

2.3 Adaptive Q Fusion at Decision Time

At decision time, since fluid is the only treatment dimension shared by both agents, BIFOCAL fuses their Q-values along this dimension using the same risk–conflict mechanism as in credit routing, plus an additional low-conflict gate: the decision weights $\mathbf{w}_t^{\mathrm{dec}}$ equal the routing weights $\mathbf{w}_t$

Model	Off-Policy Evaluation ↑					Support	Treatment Discrepancy ↓			Action Consistency ↑	
	WIS	WISt	WISb	WISbt	DR	ESSt	RMSE_{IV}	RMSE_{UF}	$\text{RMSE}_{\text{Vaso}}$	PF1	SF1
(a) Comorbidity dataset											
Clinician	76.047	76.047	76.026	76.017	76.047	6392	–	–	–	–	–
Our Model	**87.700**	**76.443**	**87.287**	**76.437**	**76.230**	4174	215.630	48.381	0.126	**0.663**	**0.628**
IQL*	66.243	75.947	38.469	75.928	68.557	4306	223.411	**29.294**	**0.114**	0.646	0.618
VDN*	82.617	65.988	82.949	65.961	28.062	82	397.763	63.168	0.146	0.050	0.049
QMIX*	83.324	72.539	67.914	72.597	66.051	1492	221.608	33.174	0.151	0.482	0.459
CFCQL◇	83.488	76.099	84.592	76.105	74.643	4124	**170.681**	58.262	0.122	0.652	0.622
B3C◇	86.726	72.605	47.490	72.632	70.771	3177	292.222	32.370	0.132	0.594	0.567
(b) AKI-only dataset											
Clinician	72.422	72.422	72.444	72.528	72.422	3386	–	–	–	–	–
Our Model	**84.294**	**74.434**	**82.514**	**74.550**	**77.561**	2839	**119.391**	38.389	0.142	**0.701**	**0.701**
IQL*	74.724	72.537	77.947	72.667	75.524	2489	261.297	24.615	**0.124**	0.650	0.650
VDN*	75.942	68.953	44.378	69.327	24.640	9	483.283	73.898	0.145	0.067	0.068
QMIX*	84.289	68.777	81.634	68.805	44.387	625	264.205	27.513	0.134	0.334	0.335
CFCQL◇	79.442	73.732	80.002	73.830	74.049	2507	158.099	31.140	0.130	0.642	0.640
B3C◇	79.858	72.827	79.740	72.894	76.217	2541	273.726	**19.821**	0.128	0.637	0.638
(c) Sepsis-only dataset											
Clinician	56.087	56.087	56.064	56.243	56.087	4580	–	–	–	–	–
Our Model	82.617	**56.277**	**82.190**	**56.424**	**61.933**	3609	**149.830**	10.263	0.135	**0.716**	**0.660**
IQL*	**85.146**	53.615	28.566	53.695	53.597	3395	257.191	**6.597**	**0.125**	0.652	0.600
VDN*	82.617	50.776	63.993	50.815	46.400	1341	228.419	26.670	0.137	0.276	0.266
QMIX*	82.617	38.171	61.090	37.517	32.179	116	362.415	106.230	0.174	0.173	0.162
CFCQL◇	83.186	-21.507	33.559	-23.440	26.584	8	514.762	130.754	0.227	0.013	0.013
B3C◇	-82.607	52.942	-61.002	53.056	58.406	3463	253.290	22.975	0.133	0.688	0.637

Table 1: **Baseline comparison across patient cohorts in MIMIC-IV.** Results are shown for **(a)** Sepsis–AKI comorbidity, **(b)** AKI-only, and **(c)** Sepsis-only cohorts. **Bold** values indicate the best result, underlined values indicate the second-best result, and dashes indicate metrics excluded from ranking because they are not applicable. ESSt is reported as the effective sample size. * denotes standard value-based MARL baselines, and ◇ denotes offline/conservative RL baselines.

from Sec. 2.2 when $C_t \geq c_0$ ($c_0 = 0.5$), and equal $(1/2, 1/2)$ otherwise. Thus, low-conflict states retain equal disease influence over fluid selection, whereas high-conflict states prioritize the disease with greater predicted deterioration risk. For each $a_{\text{F}} \in \mathcal{A}_{\text{F}}$, the AKI head maximizes $Q^{\text{AKI}}(s_t^{\text{AKI}}, (a_{\text{F}}, a_{\text{UF}}))$ over $a_{\text{UF}} \in \mathcal{A}_{\text{UF}}$ and denotes the resulting value by $Q_t^{\text{AKI,F}}(a_{\text{F}})$. Similarly, the Sepsis head maximizes $Q^{\text{SEP}}(s_t^{\text{SEP}}, (a_{\text{F}}, a_{\text{Vaso}}))$ over $a_{\text{Vaso}} \in \mathcal{A}_{\text{Vaso}}$ and denotes the result by $Q_t^{\text{SEP,F}}(a_{\text{F}})$.

BIFOCAL then computes the fluid-specific fused score and selects the coordinated fluid level as

$$Q^{\text{F}}_{\text{fuse}}(\mathbf{s}_t, a_{\text{F}}) = \sum_{d \in \{\text{AKI,SEP}\}} w_t^{\text{dec},d}\, Q_t^{d,\text{F}}(a_{\text{F}}), \quad a^*_{t,\text{F}} = \arg\max_{a_{\text{F}} \in \mathcal{A}_{\text{F}}} Q^{\text{F}}_{\text{fuse}}(\mathbf{s}_t, a_{\text{F}}). \tag{3}$$

Given the selected fluid level, the remaining disease-specific components are selected from their corresponding value heads: $a^*_{t,\text{UF}} = \arg\max_{a_{\text{UF}} \in \mathcal{A}_{\text{UF}}} Q^{\text{AKI}}(s_t^{\text{AKI}}, (a^*_{t,\text{F}}, a_{\text{UF}}))$ and $a^*_{t,\text{Vaso}} = \arg\max_{a_{\text{Vaso}} \in \mathcal{A}_{\text{Vaso}}} Q^{\text{SEP}}(s_t^{\text{SEP}}, (a^*_{t,\text{F}}, a_{\text{Vaso}}))$. The final executable treatment is $\mathbf{a}^*_t = (a^*_{t,\text{F}}, a^*_{t,\text{UF}}, a^*_{t,\text{Vaso}})$. Thus, $Q^{\text{F}}_{\text{fuse}}$ is a decision-stage coordination score defined only over the shared fluid dimension, rather than a separately learned full joint-action Q-function. The pseudocode of the decision-time procedure is given in Algorithm 2 (Appendix B.3).

Eq. 3 introduces an operator mismatch we state rather than elide: the TD target bootstraps with the unweighted $\max_{\mathbf{a}'} Q_{\text{VDN}}$, whereas deployment uses a weighted score under a hard gate at c_0. Where the two argmaxes differ the learned values do not satisfy the Bellman equation for the executed policy, and reusing $\boldsymbol{\rho}_t, C_t$ at both stages does not repair this — signal reuse, action-rule agreement, and Bellman consistency are three distinct properties. We therefore measure the gap: additive-greedy and fused-greedy joint actions coincide on 99.661% of test states, differing on 0.339% almost entirely under high conflict, so the executed policy is a small, conflict-localized perturbation of the one the values were fit for (Appendix Fig. 11).

Cohort	Model	WIS	WISt	WISb	WISbt	DR	ESSt
AKI	Our Model	**84.294**	**74.434**	**82.514**	**74.550**	77.561	2839
AKI	HRRT	74.717	71.787	76.555	71.739	**81.428**	2120
Sepsis	Our Model	**82.617**	**56.277**	82.190	**56.424**	**61.933**	3609
Sepsis	WD3QNE	**82.617**	2.683	**82.606**	2.354	-5.018	372

Table 2: **Comparison with single-disease baselines.** We also compare our model with single-disease baselines on the corresponding AKI-only and Sepsis-only cohorts.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

We evaluate BIFOCAL on Sepsis–AKI, AKI-only, and Sepsis-only cohorts from MIMIC-III, MIMIC-IV, and eICU in the U.S. (Johnson et al., 2016; 2023; Pollard et al., 2018), and a private cohort in China, totaling **87,566 trajectories** and **1,285,977 4-h decision points**. The comorbidity cohort tests BIFOCAL's ability to coordinate competing disease-oriented preferences, whereas the single-disease cohorts verify that this coordination does not compromise disease-specific decision making and assess applicability across the full disease spectrum.

We use a broader evaluation framework than prior studies to evaluate the model (Komorowski et al., 2018; Wu et al., 2023; Luo et al., 2024). **Off-policy evaluation** (higher is better) includes WIS (Tokdar & Kass, 2010), its truncated/bootstrapped variants (WISt, WISb with $B = 100$, WISbt), and DR (Jiang & Li, 2016). ESSt measures the effective sample size of truncated weights and diagnoses weight degeneracy (higher is better) (Martino et al., 2017). **Treatment discrepancy** uses RMSE for fluid, ultrafiltration, and vasopressor decisions (RMSE_{IV}, RMSE_{UF}, $\text{RMSE}_{\text{Vaso}}$) (Wang et al., 2018; Huang et al., 2022). **Action consistency** uses patient- and step-level F1 scores (PF1, SF1; higher is better) (Powers, 2020; Luo et al., 2024). Cohort construction, preprocessing, and train/test splits are detailed in Appendix C.3 (Table 6); baseline adaptation and hyperparameters in Appendix C.4; and the behavior model, truncation rule, and doubly robust nuisance fitting in Appendix C.5. Seeds and computing environment are reported in Appendices C.1–C.2.

We compare BIFOCAL with **7** RL baselines in two families. **(1) Multi-agent RL:** Independent Q-Learning (Tan, 1993), VDN (Sunehag et al., 2018), and QMIX (Rashid et al., 2018) are standard value-based MARL methods, while CFCQL (Shao et al., 2023) and B3C (Kim & Sycara, 2026) are offline/conservative MARL methods; together these five test whether existing MARL methods can coordinate disease-specific agents without the specialized credit routing. **(2) Single-disease methods:** HRRT (Xu et al., 2026) and WD3QNE (Wu et al., 2023) were specifically developed for single-disease treatment optimization; we therefore compare the model with each method in its corresponding disease-only cohort. Clinician behavior is additionally reported as a *reference policy*.

3.2 OVERALL EFFECTIVENESS

Table 1 reports results across three MIMIC-IV cohorts. On Sepsis–AKI, BIFOCAL ranks first on all five OPE metrics (e.g., WIS $= 87.70$), including WISt and WISbt, indicating higher estimated long-term value than clinician and other methods. Its ESSt remains around or above $3{,}000$, supporting reliable importance-weighted estimates. This advantage generalizes to additional datasets (Fig. 3e), where BIFOCAL outperforms the clinician policy on MIMIC-III (WISb 88.45 vs. 66.19), the private cohort (71.74 vs. 70.06), and eICU (50.09 vs. 47.32), demonstrating robustness across clinical settings. Appendix Fig. 4 further shows that this estimated value advantage is stable across bootstrap resampling, candidate checkpoints, and OPE metrics.

Standard MARL baselines confirm that value mixing alone is insufficient. VDN performs substantially worse (WISt $= 65.988$), and QMIX's nonlinear mixer offers no comparable gain, so neither decomposition nor mixer flexibility explains BIFOCAL's advantage. This advantage over VDN grows even larger in severe and high-shock subgroups, where coordination matters most (Appendix Fig. 8). BIFOCAL also outperforms CFCQL, which applies fixed conservative penalties across patient states. This result highlights the benefit of adapting agent-level weighting to each

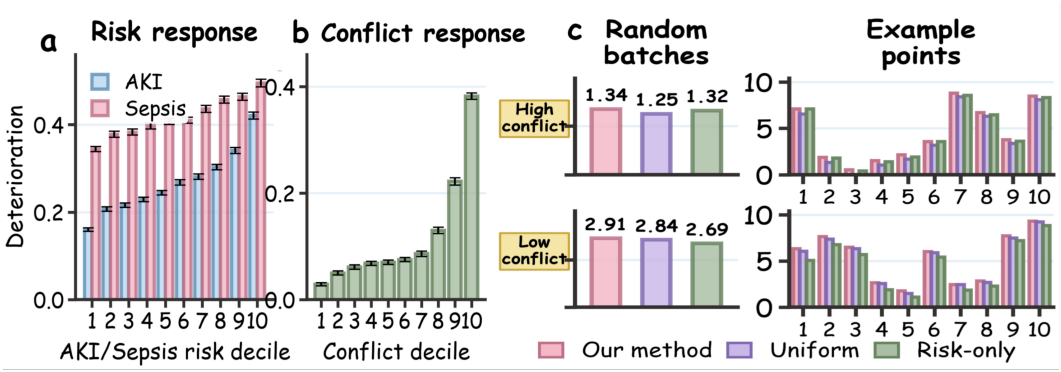


Figure 2: **Mechanism validation of risk–conflict credit routing.** **a,** Predicted AKI/Sepsis risk deciles track future deterioration. **b,** Predicted conflict deciles track JS divergence between agents' fluid preferences. **c,** One-step TD-error reduction for BIFOCAL, Uniform, and Risk-only routing under high- vs. low-conflict states; larger values indicate updates closer to TD targets. Error bars in **a,b** denote standard errors.

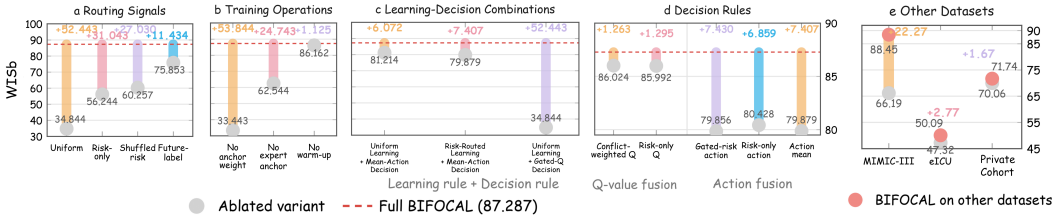


Figure 3: **Ablation analysis and cross-dataset evaluation.** **a,** Routing variants. **b,** Training protection variants. **c,** Learning–decision rule pairs. **d,** Q-space and action-space decision variants. **e,** WISb of BIFOCAL and the clinician policy on MIMIC-III, eICU, and a private cohort from China; annotations report BIFOCAL's difference from the clinician policy. The dashed line in **a–d** denotes full BIFOCAL on MIMIC-IV. Higher WISb indicates better performance.

patient's risk and treatment conflict. Overall, BIFOCAL outperforms all five comorbid-cohort baselines (IQL, VDN, QMIX, CFCQL, and B3C) on all five off-policy metrics. Notably, BIFOCAL also surpasses disease-specific models on most OPE metrics in both AKI-only and Sepsis-only cohorts (Table 2), suggesting that one comorbidity-aware model can handle both isolated and concurrent diseases, reducing the need for separate disease-specific policies and broadening its clinical utility.

Although BIFOCAL does not achieve the lowest RMSE for every treatment (Table 1), it attains the highest PF1 and SF1 across cohorts, indicating the most consistent categorical clinician agreement; the corresponding action distributions are shown in Appendix Fig. 5. This suggests that BIFOCAL remains clinically grounded while selectively identifying potentially higher-value strategies, consistent with its higher policy values.

3.3 Risk–Conflict Credit Routing

We first validate BIFOCAL's routing signals. In Fig. 2a,b, higher predicted risk precedes greater disease-specific deterioration, while higher predicted conflict reflects greater fluid-preference disagreement. Thus, risk accurately identifies which disease to prioritize, and conflict determines whether unequal routing is needed. Representative high-conflict states and their case-level interpretation are shown in Appendix Figs. 6 and 7. Panel c measures each TD update's immediate effectiveness by the one-step reduction in absolute TD error. Under high conflict, BIFOCAL achieves the largest reduction (1.34 vs. 1.32 for Risk-only and 1.25 for Uniform), showing that risk-directed routing improves updates when treatment preferences diverge. Under low conflict, it remains close to Uniform and exceeds Risk-only, indicating that conflict gating avoids unnecessary asymmetric updates when preferences align (Appendix Fig. 7). Controlled Multi-agent Particle Environment

(MPE) experiments further show improved credit allocation beyond the clinical setting, suggesting that the mechanism is not ICU-specific; because the correct credit split there is defined by construction rather than estimated, this isolates the routing rule from the clinical estimation pipeline (Appendix Fig. 10). Overall, risk determines which disease receives greater learning emphasis, while conflict determines when unequal routing is needed.

3.4 Ablation analysis

Value-Function Optimization: Credit Routing and Pretrained Value Preservation. **Fig. 3** summarizes ablations of three key components: credit routing and knowledge preservation during training, consistency between learning and decision making, and adaptive coordination at decision time. **Fig. 3a** directly evaluates BIFOCAL's **value-function optimization** mechanism by keeping the VDN backbone fixed and varying only the AKI/Sepsis learning weights. Uniform removes risk and conflict, Risk-only applies risk priority in all states, Shuffled-risk uses mismatched risks, and Future-label replaces predicted risks with ground-truth future-deterioration labels. These variants reduce WISb from 87.287 to 34.844–75.853. Even Future-label remains below BIFOCAL (75.853 vs. 87.287), showing that knowing which disease will worsen is insufficient: patient-matched risk must set the priority, while current conflict determines when unequal routing is needed. **Fig. 3b** examines preservation of pretrained branch behavior during joint training, removing the expert anchor or making its strength independent of conflict lowers WISb, whereas removing warm-up has little effect (86.162). Thus, effective comorbid learning requires both risk–conflict routing and adaptive preservation of pretrained expertise.

Consistency Between Learning and Decision Rules. **Fig. 3c** tests whether risk–conflict priority should be applied at both the learning stage and the decision stage, using a 2×2 design. During learning, *Uniform* gives equal credit to both agents, while *Routed* applies risk–conflict weights to their TD updates. At decision time, *Mean-action* simply averages the two agents' chosen actions, while *Gated-Q* uses the same risk–conflict weights to fuse their Q-values before choosing an action. Applying *Routed* at both stages performs best (WISb $= 87.287$); using *Uniform* at both stages performs worse (81.214); and mixing the two stages performs worst of all, dropping to as low as 34.844. This shows that the priority signal must be applied consistently across learning and decision-making, not just one or the other.

Decision-Level Coordination Rules. Fig. 3d compares BIFOCAL's full adaptive Shared-Fluid Action Fusion (WISb 87.287) with alternative decision rules. Q-space alternatives reduce WISb by 1.263–1.295, action-space alternatives reduce it by 6.859–7.430. These results validate BIFOCAL's decision design, which coordinates disease-specific Q-values before action selection and uses treatment conflict to control the influence of risk-based priority. Additional action-level analyses show that adaptive routing shifts shared-fluid decisions toward the preference of the higher-priority disease, confirming that the predicted disease priority is reflected in the final treatment (Appendix Fig. 11). The full routing, training-operation, and decision-rule ablations with all five estimators are reported in Appendix Tables 3, 4, and 5.

4 Conclusion and Discussion

We introduced BIFOCAL, an offline MARL framework for comorbid treatment coordination. It routes disease-specific TD updates using predicted future risk and current treatment conflict and applies the same risk–conflict principle at decision time. BIFOCAL outperforms 5 general-purpose RL baselines in the comorbid cohort and 2 disease-specific methods on most OPE metrics in disease-only cohorts. Ablations show that uniform or risk-only routing and inconsistent learning–decision rules do not reproduce these gains. A controlled particle-game study, where the correct credit split is defined by construction, isolates the direction-plus-activation structure from the clinical estimation pipeline (Appendix Fig. 10); a physician assessment appears in Appendix C.6. Offline estimates do not establish clinical benefit. Future work should extend the framework to broader multimorbidity.

AI USE STATEMENT

Generative AI tools provided limited assistance with manuscript writing and editing, development of portions of the code, and figure preparation. The authors take full responsibility for the final manuscript, code, figures, analyses, and scientific claims, including all AI-assisted content.

ETHICS STATEMENT

This work studies retrospective ICU records from MIMIC-III, MIMIC-IV, eICU, and a private clinical cohort. Patient-level data must be handled under the applicable data-use agreements and institutional permissions, including those governing the private cohort, with appropriate protection of patient confidentiality. The physician evaluation used de-identified MIMIC-III cases, involved consenting physicians, and did not affect clinical care; its design and ethics considerations are described in Appendix C.6. Off-policy estimates and physician ratings do not establish causal treatment benefit, prospective effectiveness, or clinical safety. Historical treatment practices, incomplete observations, and differences in cohort composition may introduce confounding and bias. Performance across datasets and favorable physician ratings do not establish equitable performance across patient groups. Clinical use would require prospective validation, assessment of subgroup performance and distribution shift, and clinician oversight.

REPRODUCIBILITY STATEMENT

The source code for BIFOCAL is included in the supplementary material. Sections 2.1, 2.2, and 2.3 specify the disease-conditioned value branches, credit-routing mechanism, and decision-time fusion rule. Algorithms 1 and 2 provide the training and decision procedures, while Appendices B.4–B.5 describe deterioration-target construction and risk-module pretraining. Appendices C.1–C.4 document randomness control, computing infrastructure, cohort construction, preprocessing, data splits, and model and baseline implementation settings. Appendix C.5 specifies the off-policy and action-consistency evaluation procedures, including bootstrap estimation and effective sample size. Appendix C.6 details physician participation, case sampling, the two-stage evaluation procedure, and the rating instrument.

REFERENCES

Axel Abels, Diederik M. Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic weights in multi-objective deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 11–20. PMLR, 2019.

Josée Bouchard, Sharon B. Soroko, Glenn M. Chertow, Jonathan Himmelfarb, T. Alp Ikizler, Emil P. Paganini, and Ravindra L. Mehta. Fluid accumulation, survival and recovery of kidney function in critically ill patients with acute kidney injury. *Kidney International*, 76(4):422–427, 2009. doi: 10.1038/ki.2009.159.

Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M. Coopersmith, Craig French, Flavia R. Machado, Lauralyn McIntyre, Marlies Ostermann, Hallie C. Prescott, Christa Schorr, Steven Simpson, W. Joost Wiersinga, Fayez Alshamsi, Derek C. Angus, Yaseen Arabi, Luciano Azevedo, Richard Beale, Greg Beilman, Emilie Belley-Cote, Lisa Burry, Maurizio Cecconi, John Centofanti, Angel Coz Yataco, Jan De Waele, R. Phillip Dellinger, Kent Doi, Bin Du, Elisa Estenssoro, Ricard Ferrer, Charles Gomersall, Carol Hodgson, Morten Hylander Møller, Theodore Iwashyna, Shevin Jacob, Ruth Kleinpell, Michael Klompas, Younsuck Koh, Anand Kumar, Arthur Kwizera, Suzana Lobo, Henry Masur, Steven McGloughlin, Sangeeta Mehta, Yatin Mehta, Mervyn Mer, Mark Nunnally, Simon Oczkowski, Tiffany Osborn, Elizabeth Papathanassoglou, Anders Perner, Michael Puskarich, Jason Roberts, William Schweickert, Maureen Seckel, Jonathan Sevransky, Charles L. Sprung, Tobias Welte, Janice Zimmerman, and Mitchell Levy. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021. *Intensive Care Medicine*, 47(11):1181–1247, 2021. doi: 10.1007/s00134-021-06506-y.

Carolin Fleischmann-Struzek, Lisa Mellhammar, Norman Rose, Alessandro Cassini, Kristina E. Rudd, Peter Schlattmann, Benedetta Allegranzi, and Konrad Reinhart. Incidence and mortality of hospital- and ICU-treated sepsis: results from an updated and expanded systematic review and meta-analysis. *Intensive Care Medicine*, 46(8):1552–1562, 2020. doi: 10.1007/s00134-020-06151-x.

Yong Huang, Rui Cao, and Amir Rahmani. Reinforcement learning for sepsis treatment: A continuous action space solution. In *Proceedings of the Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 631–647. PMLR, 2022.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661. PMLR, 2016.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x. URL https://doi.org/10.1038/s41597-022-01899-x.

Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney International Supplements*, 2(1):1–138, 2012.

Woojun Kim and Katia P. Sycara. B3C: A minimalist approach to offline multi-agent reinforcement learning. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1755–1763. International Foundation for Autonomous Agents and Multiagent Systems, 2026. doi: 10.65109/UOYL1486.

Matthieu Komorowski, Leo Anthony 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.

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

Zhiyao Luo, Yangchen Pan, Peter Watkinson, and Tingting Zhu. Position: Reinforcement learning in dynamic treatment regimes needs critical reexamination. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 33432–33465. PMLR, 2024. URL https://proceedings.mlr.press/v235/luo24f.html.

Luca Martino, Víctor Elvira, and Francisco Louzada. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017. doi: 10.1016/j.sigpro.2016.08.025.

David J. Middleton, Toby O. Smith, Rachel Bedford, Mark Neilly, and Phyo K. Myint. Shock index predicts outcome in patients with suspected sepsis or community-acquired pneumonia: A systematic review. *Journal of Clinical Medicine*, 8(8):1144, 2019. doi: 10.3390/jcm8081144.

Didier Payen, Anne Cornélie de Pont, Yasser Sakr, Claudia Spies, Konrad Reinhart, and Jean-Louis Vincent. A positive fluid balance is associated with a worse outcome in patients with acute renal failure. *Critical Care*, 12(3):R74, 2008. doi: 10.1186/cc6916.

Arne Peine, Ahmed Hallawa, Johannes Bickenbach, Guido Dartmann, Lejla Begic Fazlic, Anke Schmeink, Gerd Ascheid, Christoph Thiemermann, Andreas Schuppert, Ryan Kindle, Leo Celi, Gernot Marx, and Lukas Martin. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *npj Digital Medicine*,

4(1):32, 2021. doi: 10.1038/s41746-021-00388-6. URL https://doi.org/10.1038/s41746-021-00388-6.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018. doi: 10.1038/sdata.2018.178. URL https://doi.org/10.1038/sdata.2018.178.

David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, 2020. URL https://arxiv.org/abs/2010.16061.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4295–4304. PMLR, 2018. URL https://proceedings.mlr.press/v80/rashid18a.html.

Kristina E. Rudd, Sarah Charlotte Johnson, Kareha M. Agesa, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet*, 395(10219):200–211, 2020. doi: 10.1016/S0140-6736(19)32989-7.

Jianzhun Shao, Yun Qu, Chen Chen, Hongchang Zhang, and Xiangyang Ji. Counterfactual conservative Q learning for offline multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77290–77312. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f3f2ff9579ba6deeb89caa2fe1f0b99c-Abstract-Conference.html.

Mervyn Singer, Clifford S. Deutschman, Christopher W. Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8): 801–810, 2016. doi: 10.1001/jama.2016.0287.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2085–2087. International Foundation for Autonomous Agents and Multiagent Systems, 2018. doi: 10.65109/JSRC7365.

Tomonori Takeuchi, Alexander H. Flannery, Lucas J. Liu, Lama Ghazi, Augusto Cama-Olivares, Kiyohide Fushimi, Jin Chen, Sarah C. Huen, Ashita J. Tolwani, and Javier A. Neyra. Epidemiology of sepsis-associated acute kidney injury in the ICU with contemporary consensus definitions. *Critical Care*, 29(1):128, 2025. doi: 10.1186/s13054-025-05351-5.

Ming Tan. Multi-agent reinforcement learning: Independent versus cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337. Morgan Kaufmann, 1993. doi: 10.1016/B978-1-55860-307-3.50049-6.

Surya T Tokdar and Robert E Kass. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010. doi: 10.1002/wics.56.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 2447–2456, New York, NY, USA, 2018. Association for Computing Machinery. doi: 10.1145/3219819.3219961. URL https://doi.org/10.1145/3219819.3219961.

Xiangsen Wang, Haoran Xu, Yinan Zheng, and Xianyuan Zhan. Offline multi-agent reinforcement learning with implicit global-to-local value regularization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52413–52429. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a46c84276e3a4249ab7dbf3e069baf7f-Abstract-Conference.html.

Ziyan Wang, Yali Du, Yudi Zhang, Meng Fang, and Biwei Huang. MACCA: Offline multi-agent reinforcement learning with causal credit assignment. *Transactions on Machine Learning Research*, June 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=gwUOzI4DuV.

XiaoDan Wu, RuiChang Li, Zhen He, TianZhi Yu, and ChangQing Cheng. A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. *npj Digital Medicine*, 6(1):15, 2023. doi: 10.1038/s41746-023-00755-5.

Qianyi Xu, Feng Wu, Zi Yi Christopher Thong, Mark Sen Liang Goh, Pengpeng Chen, Jie Yang, Chen Huang, Zhongheng Zhang, Yucai Hong, Kay Choong See, et al. Hrrt: hierarchical reinforcement learning for renal replacement therapy decision support. *npj Digital Medicine*, 2026.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14610–14621, 2019.

Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10299–10312. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/550a141f12de6341fba65b0ad0433500-Abstract.html.

Kristine Zhang, Henry Wang, Jianzhun Du, Brian Chu, Aldo Robles Arévalo, Ryan Kindle, Leo Anthony Celi, and Finale Doshi-Velez. An interpretable RL framework for pre-deployment modeling in ICU hypotension management. *npj Digital Medicine*, 5(1):173, 2022. doi: 10.1038/s41746-022-00708-4.

Yihe Zhou, Yuxuan Zheng, Yue Hu, Kaixuan Chen, Tongya Zheng, Jie Song, Mingli Song, and Shunyu Liu. Cooperative policy agreement: Learning diverse policy for offline MARL. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23018–23026, 2025. doi: 10.1609/aaai.v39i21.34465. URL https://ojs.aaai.org/index.php/AAAI/article/view/34465.

Overview This appendix provides supporting analyses, methodological details, and reproducibility information. Appendix A reports additional empirical analyses and validation results; Appendix B gives supplementary method details for the comorbid multi-agent formulation, training and decision procedures, and risk module; Appendix C summarizes seeds, computing environment, datasets, baselines, off-policy evaluation, and clinician-evaluation considerations.

A Additional Analyses and Validation

A.1 Stability of Estimated Policy Value

This section evaluates whether BIFOCAL's OPE-estimated policy value advantage is stable, and examines whether the main results could be driven by a specific bootstrap setting, a single high-performing candidate checkpoint, or one particular OPE metric. Fig. 4 provides three complementary analyses.

First, Fig. 4a evaluates the sensitivity of the estimated value to bootstrap settings. We vary the bootstrap resampling scale and compare the WISbt estimated return across methods. BIFOCAL maintains the highest WISbt return under all resampling settings, with no substantial change in its relative advantage. This suggests that its OPE performance does not depend on a particular bootstrap configuration.

Second, Fig. 4b examines stability across candidate models. As more candidate models are included, we compute the cumulative median WISbt return to test whether BIFOCAL's high estimated return is mainly driven by a single strong checkpoint. BIFOCAL remains above the clinician reference and other baselines across candidate counts, indicating that its advantage is not caused by one isolated checkpoint but persists within the candidate model set.

Finally, Fig. 4c shows the overall behavior of candidate models across multiple OPE metrics. If the result were driven by a single model or a single metric, high performance would concentrate in a few isolated model–metric combinations. Instead, the high-performance regions appear as relatively continuous and consistent patterns across candidate models and OPE metrics, suggesting that BIFOCAL's advantage is not an accidental peak on one metric.

Overall, Fig. 4 shows that BIFOCAL's estimated policy value advantage is stable across bootstrap resampling, candidate model aggregation, and multi-metric OPE evaluation. This analysis complements the main results by showing that the observed advantage is not driven by a single experimental setting, checkpoint, or OPE metric.

A.2 Model Behavior Visualization

Beyond OPE, the RMSE and categorical F1 metrics reported in the main text (Table 1) assess the policy's agreement with clinician practice. Although BIFOCAL is not closest to clinicians on every RMSE metric, its discrepancies remain comparable and it achieves the highest F1 scores, suggesting that it often preserves the clinician action category while allowing moderate intensity changes, especially through the shared fluid channel. Fig. 5 further visualizes this pattern. Our policy keeps a broadly clinician-like IV-fluid distribution, but makes localized departures by reducing high-intensity ultrafiltration and vasopressor use. In contrast, other methods such as IQL and VDN show more scattered action patterns. These results suggest that our performance gain is associated with structured local deviations rather than uniform clinician imitation or broad action-space drift.

A.3 Learning-Level Routing Analysis

Interpretability of high-conflict states The cases in Fig. 6 further show that these high-conflict states reflect interpretable treatment tension rather than numerical noise. Across the four examples, high conflict appears either when AKI and Sepsis prefer opposite fluid directions or when an aggressive action is paired with an unstable patient state.

Case-level allocation behavior Fig. 7 evaluates the final allocation. In the high-conflict case, the gate is open, so the risk difference changes the update weights toward the disease with higher predicted deterioration. In the low-conflict case, the gate is nearly closed, so the weights remain

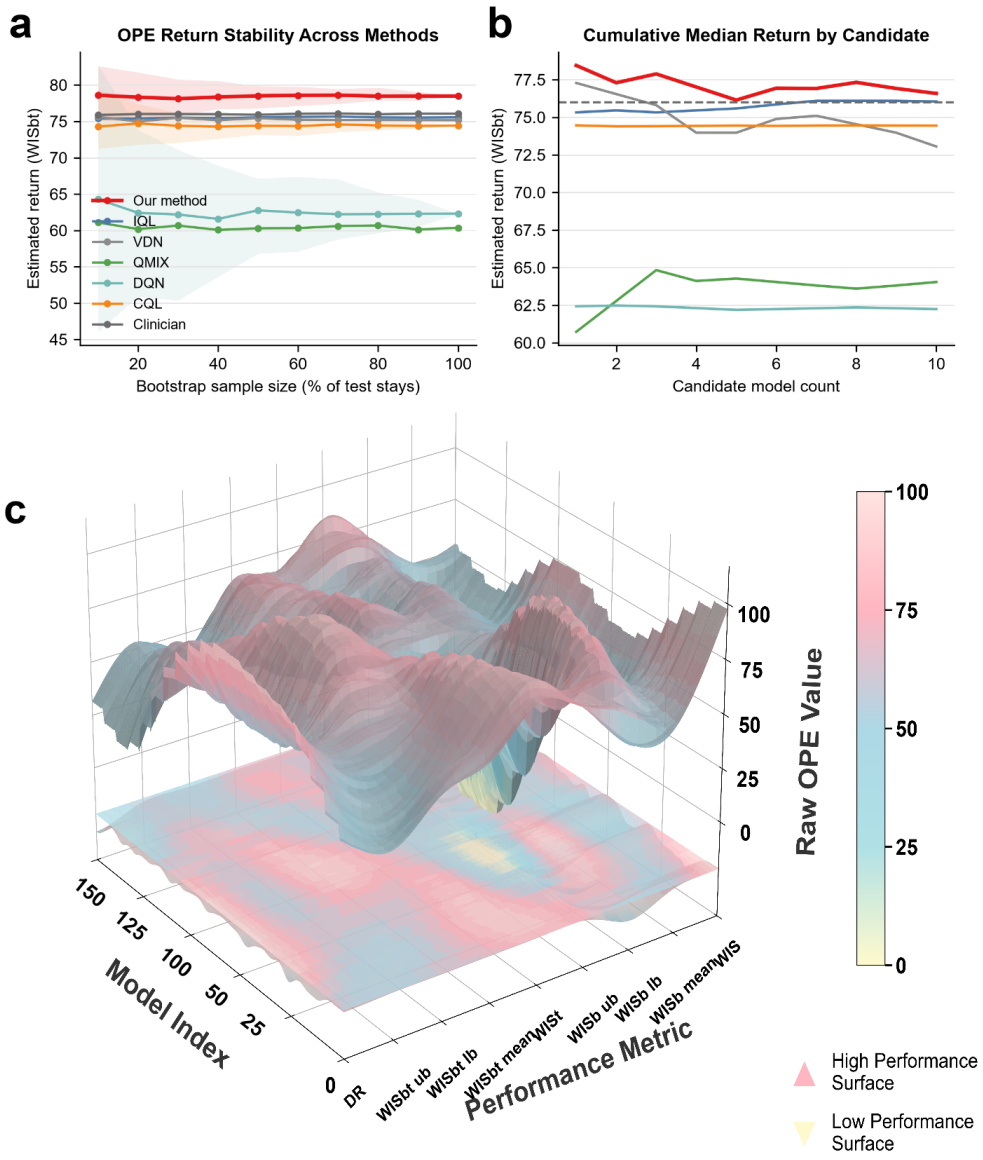


Figure 4: **Stability of estimated policy value. a,** Estimated WISbt return under different bootstrap sample sizes on the test cohort. **b,** Cumulative median return as candidate models are included, with the dashed line indicating the clinician reference. **c,** Three-dimensional OPE landscape across candidate models and evaluation metrics. The x-axis indexes candidate models, the y-axis lists OPE metrics, and the z-axis shows raw OPE values. The upper surface highlights high-performance regions, the projected plane shows contour structure, and the lower reference layer summarizes low-performance variation.

balanced even when the two risks are not exactly equal. This shows that risk does not directly control the weights at every step. It is used only when current treatment conflict makes asymmetric routing necessary.

Subgroup performance Fig. 8 examines whether routing gains remain stable across clinically different subgroups. Our method keeps high WISbt across all groups, ranging from 74.339 in high-shock states to 81.696 in mild states. The gain over VDN is small in easier groups, including mild

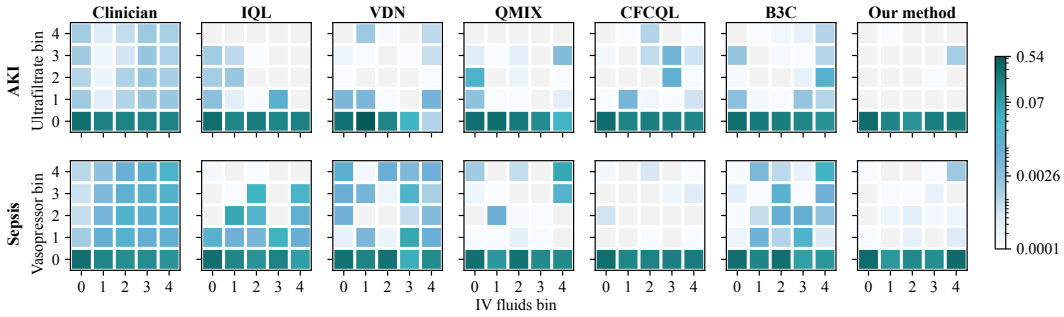


Figure 5: Policy action distributions on the test cohort. Rows show AKI and Sepsis agents; columns compare clinician actions and learned policies.

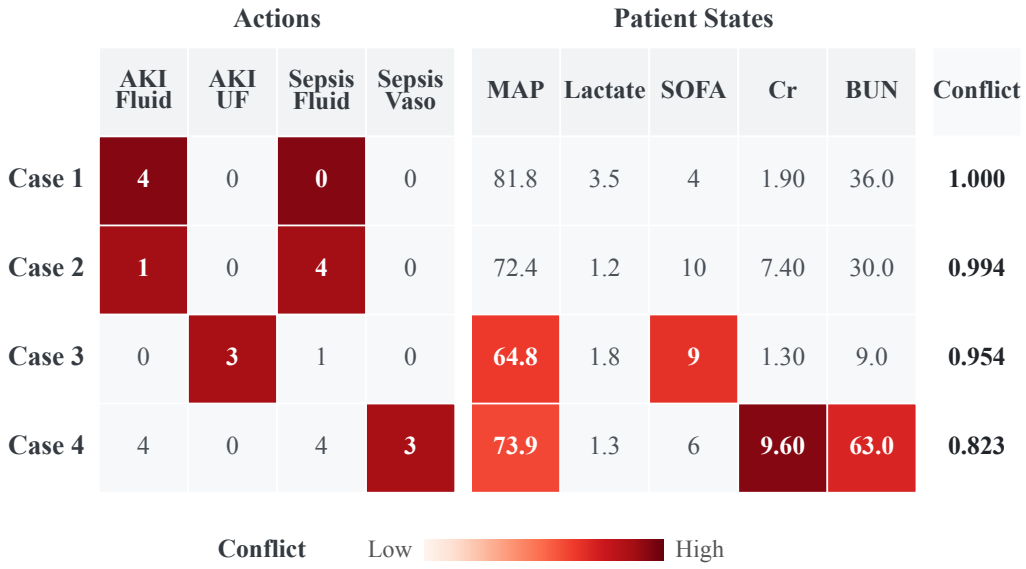


Figure 6: High-conflict case analysis. Four representative cases with high conflict scores computed by the conflict router are shown. Red cells indicate case-specific evidence supporting each conflict interpretation, with darker red denoting stronger conflict evidence.

states and low-shock states, but becomes larger in harder groups. The gain is 7.283 in severe states versus 0.843 in mild states, 5.586 in high-shock states versus 0.471 in low-shock states, and 6.505 in balanced-dominance states versus 4.188 in single-disease-dominant states. The gain is also positive in high-conflict states and in low-conflict states. This suggests that the router helps in two ways. It shifts learning weights when coordination is difficult and keeps updates stable when conflict is limited. Overall, these subgroup results support that risk–conflict routing is most useful when AKI–Sepsis coordination is harder, while remaining effective across easier states.

A.4 Additional Ablations

Routing-signal ablation Table 3 keeps the VDN backbone fixed and changes only how AKI/Sepsis learning weights are computed. Uniform routing removes both risk and conflict, Risk-only removes conflict activation, Shuffled risk breaks the match between patient state and predicted risk, and Future-label replaces graded predictions with hard future-deterioration labels. All variants reduce WISb relative to the full model, supporting that routing needs both patient-matched risk direction and current conflict gating.

Training-operation ablation Table 4 evaluates training operations used to stabilize offline joint optimization, including expert anchoring, anchor weighting, warm-up initialization, and the com-

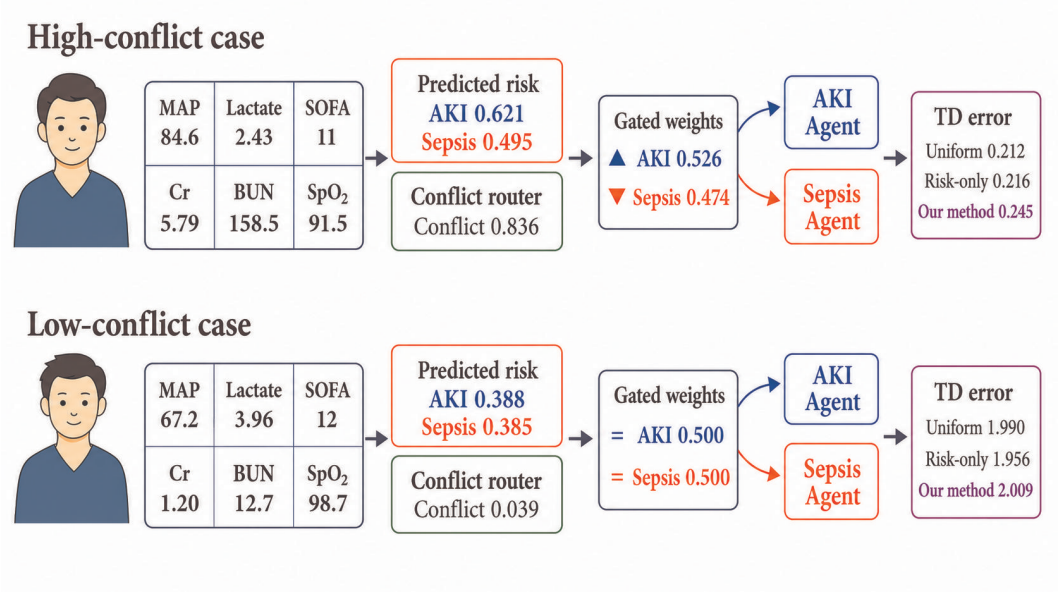


Figure 7: Case-level illustration of conflict-gated risk-aware routing. The high-conflict case activates risk-dependent update allocation, while the low-conflict case keeps the gated weights balanced despite different predicted risks. TD reduction values report the one-step reduction in absolute TD error under different credit-assignment strategies.

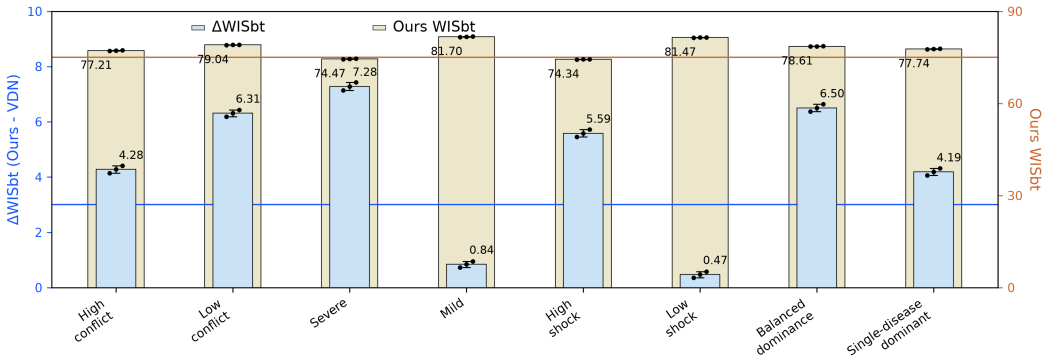


Figure 8: Subgroup analysis of routing gains. Blue bars show the gain over VDN, $\Delta\mathrm{WISbt} = \mathrm{WISbt}_{\mathrm{ours}} - \mathrm{WISbt}_{\mathrm{VDN}}$. Beige bars show our WISbt. Groups are stratified by conflict level, disease severity, shock status, and disease dominance. Error bars denote standard errors.

bined removal of routing and anchoring. Removing the expert anchor loss keeps WIS close to the full model (87.752 vs. 87.700), but substantially reduces WISb from 87.287 to 62.544, WISbt from 76.437 to 73.885, and DR from 76.230 to 69.288. This suggests that expert anchoring helps prevent excessive policy drift from independently trained disease-specific experts during joint training.

Removing the anchor weight causes a larger degradation, especially in WISb and DR. WISb decreases from 87.287 to 33.443 and DR decreases from 76.230 to 3.323. This shows that applying the anchor constraint without conflict-dependent weighting is insufficient: the model benefits from modulating the anchor loss according to action conflict. The combined removal of routing and anchoring further weakens temporal and doubly robust evaluation, and removing the warm-up stage leads to the most severe degradation in temporal metrics. Together, these results show that expert anchoring, anchor weighting, adaptive routing, and warm-up initialization all contribute to stable offline joint policy learning.

Routing Signal	WIS	WISt	WISb	WISbt	DR
Ours	87.700	**76.443**	**87.287**	**76.437**	**76.230**
Future-label routing	87.752	71.412	75.853	71.242	-0.806
Risk-only	87.752	66.802	56.244	67.011	1.986
Shuffled risk	80.041	72.904	60.257	73.003	26.381
Uniform	**87.752**	62.570	34.844	63.060	3.924

Table 3: Routing-signal ablation on the comorbidity test cohort.

Model	WISb	DR
w/o Expert Anchor Loss	62.544	69.288
w/o Conflict-Weighted Anchor Loss	33.443	3.323
w/o Dynamic Credit Routing & Anchor Loss	86.872	28.620
w/o Warm-up	86.162	12.706
Full Model	**87.287**	**76.230**

Table 4: Training-operation ablation on the comorbidity test cohort.

Decision and cross-stage ablation Table 5 reports how learning-stage routing interacts with decision-time fusion. The full conflict-gated risk-weighted Q fusion performs best. Continuous conflict-weighted and risk-only Q fusion remain close, whereas gated-risk, risk-only, and mean action fusion perform substantially worse. These results show that most of the gain comes from coordinating before action selection in Q-space, with conflict-triggered gating providing an additional benefit.

Future-risk horizon selection The future-risk module requires a predefined prediction horizon because risk prediction labels are constructed according to whether disease-specific deterioration events occur within the future window. Therefore, the horizon choice directly changes the clinical meaning of the risk labels and further affects how useful the risk signal is for downstream routing and policy learning. A very short horizon may not contain sufficiently stable information about subsequent deterioration, whereas an overly long horizon may introduce more distant factors and weaken the association between the current patient state and the future deterioration label. Therefore, we compare different future-risk horizons in terms of both risk prediction quality and downstream policy value to select the most suitable foresight window for routing.

In Fig. 9, we compare prediction horizons from 4 to 32 hours. Risk prediction quality is evaluated by the agreement between predicted risk scores and the corresponding future-window deterioration labels, using mean correlation and mean NMSE. As the horizon increases from 4 hours to 24 hours, the agreement between predicted risks and future deterioration labels improves, with increasing mean correlation and decreasing mean NMSE. This suggests that the 24-hour window provides a more stable and forward-looking supervision signal, allowing the risk module to better learn disease-specific deterioration tendency.

Downstream policy evaluation shows a consistent pattern with the risk prediction results. Although WIS varies only mildly across different horizons and remains relatively high overall, the 24-hour horizon still maintains strong overall policy value. A clearer horizon-dependent trend is observed in WISt: as the horizon increases from 4 hours to 24 hours, WISt steadily improves, whereas it decreases when the horizon is further extended to 28 or 32 hours. These results suggest that the 24-hour window not only improves risk prediction quality but also provides a more useful risk signal for downstream policy learning.

Therefore, we use 24 hours as the future-risk prediction horizon in the full model. This choice is not merely a robustness setting, but defines the temporal scope of future deterioration in the risk module. The 24-hour window provides a favorable balance between prediction-label consistency and downstream policy value, making it suitable as the forward-looking disease-risk signal for risk-conflict routing.

Learning Weights	Decision Weights	Fusion Level	WISb	DR
Risk×Conflict	Conflict-gated risk	Q-value	**87.287**	**76.230**
Risk×Conflict	Conflict-gated risk	Action	79.856	65.052
Risk×Conflict	Conflict-weighted risk	Q-value	86.024	75.767
Risk×Conflict	Risk only	Q-value	85.992	75.969
Risk×Conflict	Risk only	Action	80.428	67.698
Risk×Conflict	Action Mean	Action	79.879	65.118
Uniform	Action Mean	Action	81.214	7.224
Uniform	Conflict-gated risk	Q-value	34.844	5.162

Table 5: Decision and cross-stage ablation on the comorbidity test cohort.

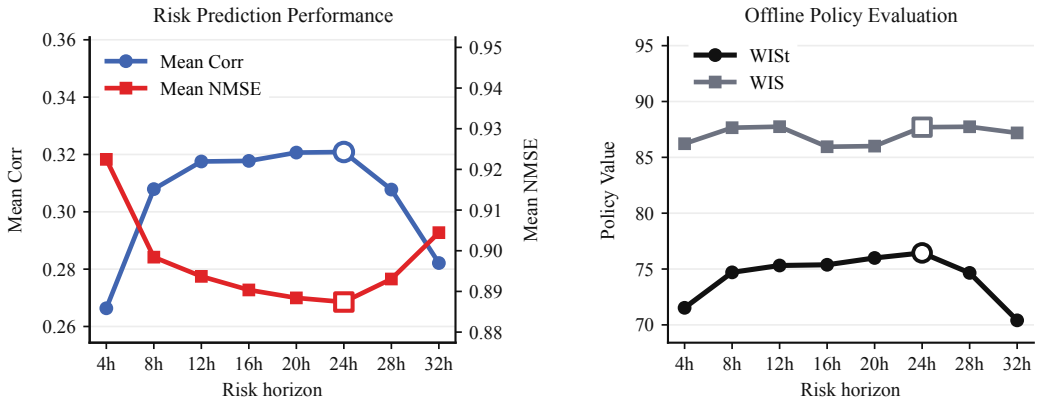


Figure 9: Future-risk horizon ablation on the comorbidity cohort. **(a)** Risk prediction quality as the prediction horizon varies from 4h to 32h, measured by mean correlation (higher is better) and mean NMSE (lower is better) between the predicted risk and observed future deterioration. **(b)** Corresponding offline policy value (WIS and WISt) when the resulting risk signal is used for routing. The 24h horizon (open markers) jointly attains the best risk prediction and the highest policy value, and is adopted throughout.

A.5 Out-of-Domain Game Validation

BIFOCAL has been validated in ICU comorbid treatment through off-policy evaluation, ablation studies, risk–conflict response analysis, and clinician evaluation. To further examine whether the proposed routing mechanism generalizes beyond the clinical data distribution, an out-of-domain MPE-style particle game is used as a complementary validation. This experiment is not designed to simulate ICU treatment. Instead, it abstracts the central routing structure of BIFOCAL: priority determines the direction of credit allocation, whereas conflict determines when asymmetric allocation should be activated.

The environment consists of two cooperative agents, one high-value prey, and one lower-value decoy. The high-value prey defines the priority direction, as the agent with better access to this target is expected to contribute more when coordination conflict arises. The decoy does not directly represent conflict; rather, it introduces an alternative target, allowing states to range from natural role splitting to redundant competition. Each agent forms a target preference between the prey and the decoy according to target distance and prey urgency. Conflict is low when the agents prefer different targets and can naturally split roles, and high when both agents prefer the same target, indicating overlapping roles or redundant competition. The conflict score further incorporates prey urgency and the priority gap, so that it reflects target-preference overlap, target importance, and responsibility asymmetry.

Three routing strategies are evaluated: **Uniform routing**, which always assigns equal credit; **Priority-only routing**, which always follows the priority direction regardless of conflict; and

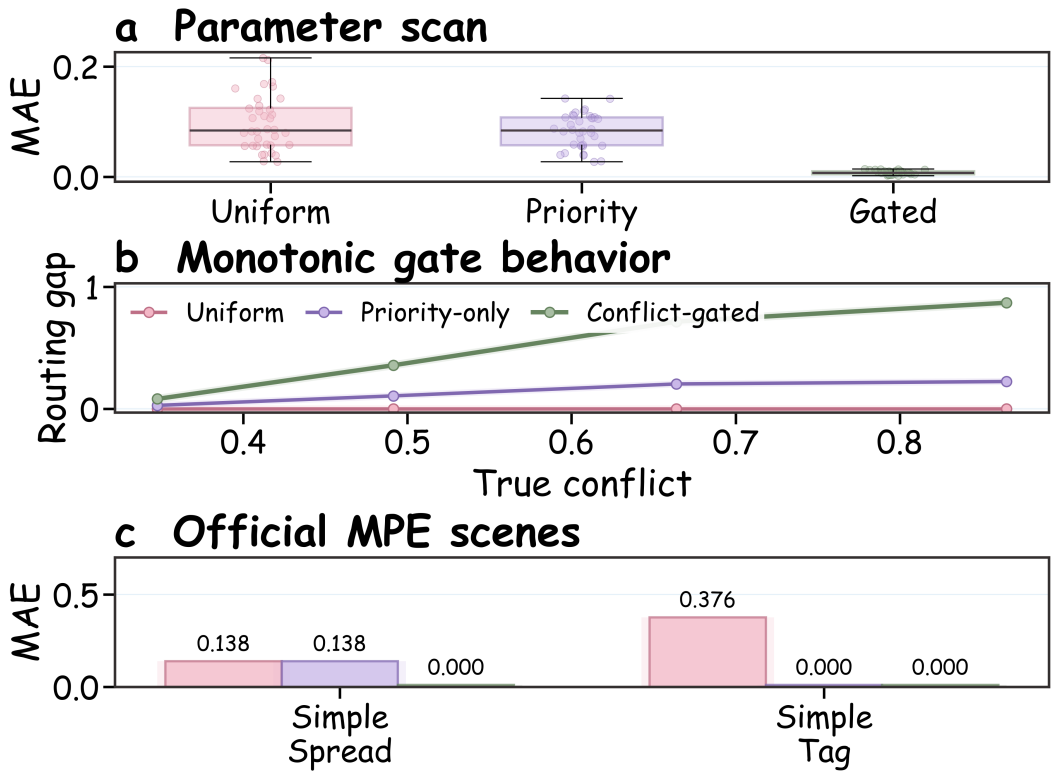


Figure 10: **Out-of-domain validation of credit routing.** MPE-style games provide controlled state-defined routing targets for how team credit should be divided between agents. Uniform assigns equal credit, Priority-only follows agent priority regardless of conflict, and Conflict-gated applies unequal credit only when agents conflict. **a,** MAE measures deviation from the target routing weights across parameterized game settings. Each dot represents one setting. **b,** The routing gap measures the credit difference between agents. It rises with state-defined conflict under Conflict-gated routing, remains zero under Uniform, and changes weakly under Priority-only. **c,** MAE on official Simple Spread and Simple Tag tests generalization to standard MPE tasks. Lower MAE is better.

Conflict-gated routing, which activates priority-based asymmetric allocation only under conflict. Fig. 10a shows that Conflict-gated routing produces the lowest routing error across parameterized game settings. Fig. 10b shows that its credit gap remains small in low-conflict states and increases as conflict becomes stronger, indicating that effective credit routing requires both priority direction and conflict-based activation. Fig. 10c further shows that Conflict-gated routing maintains low routing error on held-out MPE scenes when conflict and non-conflict states are mixed. This out-of-domain result provides complementary evidence that BIFOCAL captures a general coordination mechanism rather than only fitting the ICU data distribution.

A.6 Action-Level Directionality of Adaptive Q Fusion

We further examined whether the disease-priority weights produced by adaptive routing actually affect the final shared-fluid decision in the intended direction. We considered 7,874 high-conflict decisions for which the AKI and Sepsis Q-heads preferred different fluid levels. Keeping both Q-heads fixed, we compared BIFOCAL's adaptive fusion with a counterfactual that assigned equal weight to the two diseases. This comparison isolates the effect of adaptive routing because only the fusion weights are changed.

As shown in Fig. 11, adaptive routing changed the selected fluid in 3.19% of these decisions (95% CI: 2.78–3.62%). Although routing therefore altered only decisions near the fusion boundary, 95.62% of the changed decisions (92.89–97.93%) moved closer to the fluid preferred by the higher-weight disease head. All changes increased the centered Q-value of that disease, and exact

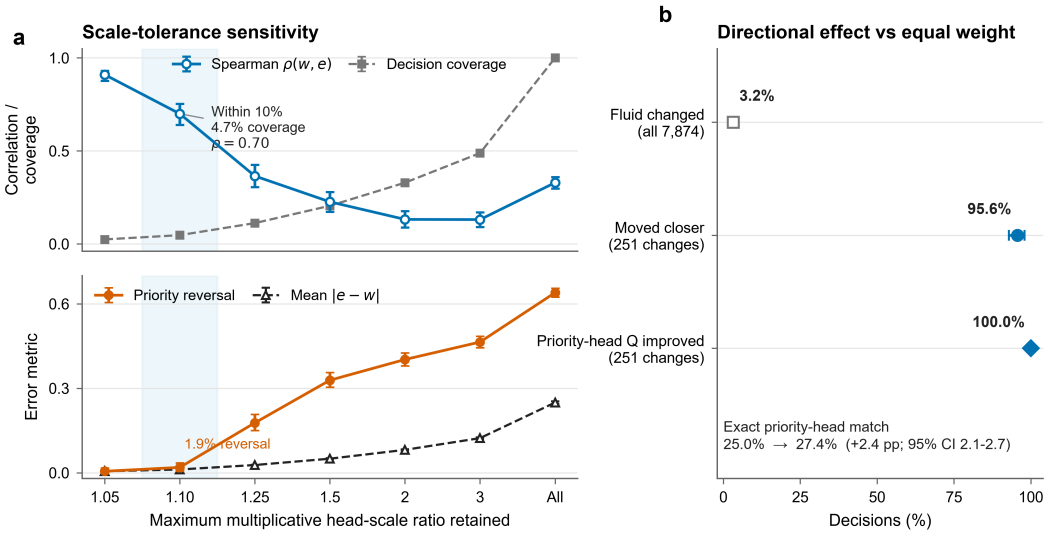


Figure 11: Action-level directionality and scale sensitivity of adaptive Q fusion. Using equal-weight fusion of the same frozen Q-heads as a counterfactual, adaptive routing changes only a small subset of high-conflict shared-fluid decisions, but these changes overwhelmingly move toward the fluid preferred by the higher-priority disease head. The scale-sensitivity analysis further shows that routing weights closely track magnitude-based Q influence when the two fluid profiles have comparable variation, while this correspondence weakens as their scales diverge. Confidence intervals are obtained by patient-level bootstrap.

agreement with its preferred fluid increased from 25.01% under equal weighting to 27.39% under adaptive routing, a net gain of 2.39 percentage points (2.07–2.75). These results show that the routing weights are reflected in the final treatment: when adaptive routing changes the shared-fluid choice, the change is overwhelmingly directed toward the disease assigned greater priority.

We additionally examined whether differences in the variation of the two fluid-specific Q profiles affect this correspondence. When the two profiles had similar magnitudes, the intended routing weights closely matched their magnitude-based influence on fusion. This agreement progressively weakened as the scale difference increased. Because this magnitude-based measure summarizes variation across fluid levels rather than the actual selected action, we treat it as a sensitivity diagnostic rather than evidence of globally aligned or identifiable Q-head scales. The controlled action-level comparison above provides the primary evidence that BIFOCAL's predicted disease priority is translated into the intended treatment direction.

B Supplementary Method Details

B.1 Problem Formulation

Clinical task and offline multi-agent MDP We address the task of dynamically coordinating treatments for patients with coexisting Sepsis and acute kidney injury (AKI). Given the sequential nature of comorbid clinical decision-making and the need to preserve disease-specific treatment considerations while optimizing a shared long-term patient outcome, we formulate this task as an offline cooperative multi-agent Markov decision process. Let $\mathcal{D}_{\text{com}} = \{\tau_i\}_{i=1}^{N}$ denote the retrospective comorbid ICU trajectory dataset, where each trajectory is observed at fixed 4-hour decision intervals. The cooperative MDP is $\mathcal{M} = (\mathcal{N}, \mathcal{S}, \{\mathcal{O}^d\}_{d\in\mathcal{N}}, \{\mathcal{A}^d\}_{d\in\mathcal{N}}, P, R, \gamma)$, where $\mathcal{N} = \{\text{AKI}, \text{SEP}\}$. Here, $\mathcal{S}$ is the aligned comorbid patient state space, $\mathcal{O}^d$ is the disease-conditioned observation space for agent d, $\mathcal{A}^d$ is its disease-specific action space, P is the transition kernel, R is the shared reward function, and γ is the discount factor. In the implementation, $\gamma = 0.99$.

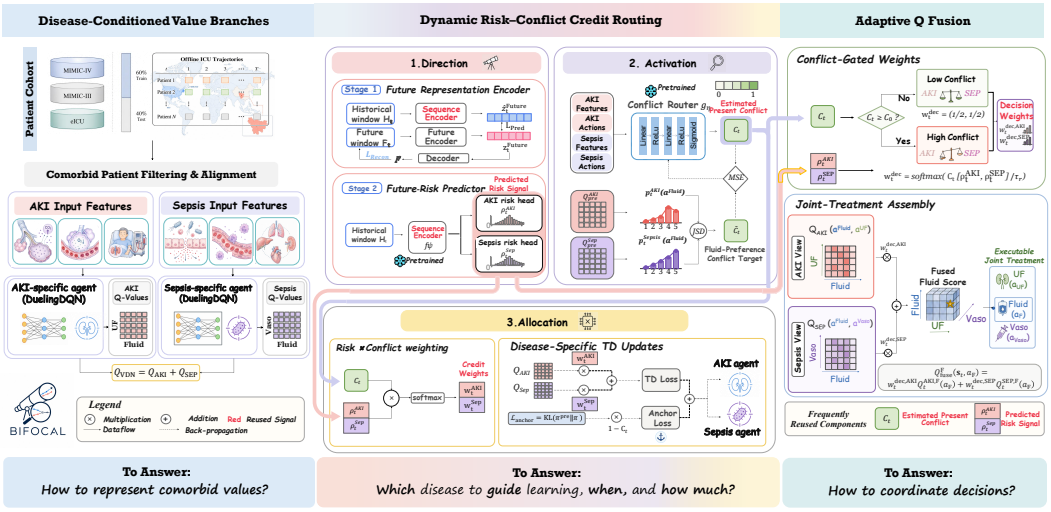


Figure 12: **Overview of BIFOCAL for Sepsis–AKI treatment coordination. Disease-Conditioned Value Branches.** AKI- and Sepsis-oriented agents learn separate Q-functions from aligned offline ICU trajectories. ***Dynamic Risk–conflict Credit Routing.*** Predicted future risk determines the direction of disease priority, current treatment conflict controls its activation, and their coupling allocates TD credit while an adaptive anchor preserves pretrained branch behavior. ***Adaptive Q Fusion.*** At decision time, a similar risk–conflict mechanism adaptively fuses the disease-specific Q-values to produce a coordinated joint treatment. This is the complete pipeline underlying the compact summary in Fig. 1 of the main text.

Agents and disease-conditioned state space The AKI agent represents the AKI-oriented treatment process, while the Sepsis agent represents the Sepsis-oriented treatment process. Both agents correspond to the same patient and the same 4-hour decision step. The global state is $\mathbf{s}_t = (s_t^{\mathrm{AKI}}, s_t^{\mathrm{SEP}})$, where $s_t^{\mathrm{AKI}} \in \mathcal{O}^{\mathrm{AKI}}$ and $s_t^{\mathrm{SEP}} \in \mathcal{O}^{\mathrm{SEP}}$ are disease-conditioned views of the same ICU state. The two views share common variables such as demographics, vital signs, ventilation indicators, laboratory values, severity scores, fluid input, urine output, and cumulative fluid balance, while retaining disease-specific features. The AKI view includes renal and RRT-related variables such as creatinine, BUN, urine output, ultrafiltration-related fields, and renal-support indicators; the Sepsis view emphasizes hemodynamic and sepsis-severity variables such as lactate, mean arterial pressure, vasopressor intensity, SOFA score, and shock index.

Disease-specific and joint action spaces Each agent has a two-dimensional discrete action space. The AKI action space is $\mathcal{A}^{\mathrm{AKI}} = \mathcal{A}^{F} \times \mathcal{A}^{UF}$, combining intravenous fluid and ultrafiltration. The Sepsis action space is $\mathcal{A}^{\mathrm{SEP}} = \mathcal{A}^{F} \times \mathcal{A}^{\mathrm{Vaso}}$, combining intravenous fluid and vasopressor intensity. Because fluid is a shared treatment administered once to the patient, the executable joint action space is $\mathcal{A}_{\mathrm{joint}} = \mathcal{A}^{F} \times \mathcal{A}^{UF} \times \mathcal{A}^{\mathrm{Vaso}}$. Thus, the joint treatment is $\mathbf{a}_t = (a_{t,F}, a_{t,UF}, a_{t,\mathrm{Vaso}})$, while the disease-specific sub-actions are $a_t^{\mathrm{AKI}} = (a_{t,F}, a_{t,UF})$ and $a_t^{\mathrm{SEP}} = (a_{t,F}, a_{t,\mathrm{Vaso}})$. The two agents therefore cannot execute two different fluid levels; every valid joint action contains one shared fluid level, one AKI-specific ultrafiltration level, and one Sepsis-specific vasopressor level.

Discretization of treatment actions Actions are constructed at the same 4-hour resolution as the state trajectories. The fluid variable is `input_step`, computed as total equivalent IV fluid administered within the interval; continuous infusions are integrated by overlap with the 4-hour window, boluses are added when their timestamps fall inside the window, and fluid amounts are converted to total equivalent volume before aggregation. In the comorbid data, the shared-fluid source is the Sepsis view by default and is written to both disease views before action binning. The vasopressor variable is `max_dose_vaso`, the maximum norepinephrine-equivalent dose observed in the interval. Norepinephrine and epinephrine are kept on the $\mu\mathrm{g/kg/min}$ scale when recorded that way and divided by an assumed 80 kg body weight when recorded as $\mu\mathrm{g/min}$; phenylephrine is multiplied by 0.45, dopamine by 0.01, and vasopressin by 5, with hourly units divided by 60.

The ultrafiltration variable is `Ultrafiltrate_Output`, represented as the 4-hour window-level ultrafiltration output measurement in the AKI view; missing action-source values are set to zero before discretization. For each treatment dimension $j \in \{F, UF, \text{Vaso}\}$, positive nonzero training-set doses are split into four nonzero bins using empirical quartiles, while the first bin corresponds to no treatment. The saved MIMIC-IV comorbid checkpoints use five levels per treatment dimension, so $|\mathcal{A}^F| = |\mathcal{A}^{UF}| = |\mathcal{A}^{\text{Vaso}}| = 5$, $|\mathcal{A}^{\text{AKI}}| = 25$, $|\mathcal{A}^{\text{SEP}}| = 25$, and $|\mathcal{A}_{\text{joint}}| = 125$.

Transition function The transition kernel $P(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$ denotes the probability that a patient moves from the current aligned comorbid state to the next 4-hour state after receiving joint treatment $\mathbf{a}_t$. This transition is influenced by both disease processes and their interaction, including the effect of shared fluid therapy on renal and hemodynamic trajectories. The transition dynamics are unknown and are not specified analytically; in the offline RL setting, they are learned only through retrospective transitions observed in $\mathcal{D}_{\text{com}}$.

Reward function Both agents optimize the same patient-level outcome reward based on the binary 90-day survival status recorded for each trajectory. The implementation uses a terminal-only reward and does not incorporate intermediate physiological shaping terms such as SOFA, lactate, or MAP. Let $Y_i \in \{0, 1\}$ denote the 90-day mortality indicator for trajectory τ_i, where $Y_i = 0$ denotes survival at 90 days and $Y_i = 1$ denotes death within the 90-day follow-up period. With $R_{\max} = 100$, all nonterminal transitions receive $r_t = 0$, while the terminal transition receives $r_t = +R_{\max}$ for 90-day survival and $r_t = -R_{\max}$ for 90-day mortality. Therefore, the implemented terminal rewards are $+100$ and -100, respectively.

Joint optimization objective via value decomposition BIFOCAL preserves two disease-specific action-value functions, $Q_{\text{AKI}}(s_t^{\text{AKI}}, a_t^{\text{AKI}})$ and $Q_{\text{SEP}}(s_t^{\text{SEP}}, a_t^{\text{SEP}})$. The AKI branch evaluates AKI-oriented fluid–ultrafiltration sub-actions, and the Sepsis branch evaluates Sepsis-oriented fluid–vasopressor sub-actions, while both optimize the same shared patient outcome reward. The VDN joint value is $Q_{\text{VDN}}(\mathbf{s}_t, \mathbf{a}_t) = Q_{\text{AKI}}(s_t^{\text{AKI}}, a_t^{\text{AKI}}) + Q_{\text{SEP}}(s_t^{\text{SEP}}, a_t^{\text{SEP}})$. Here, Q_{VDN} is the value of an executable joint treatment, and the joint action must satisfy shared-fluid consistency through $\mathbf{a}_t = (a_{t,F}, a_{t,UF}, a_{t,\text{Vaso}})$. Using target networks, the TD target is

$$y_t = r_t + \gamma\big(1 - \text{done}_t\big) \max_{\mathbf{a}' \in \mathcal{A}_{\text{joint}}} Q^-_{\text{VDN}}\big(\mathbf{s}_{t+1}, \mathbf{a}'\big)$$

, where $Q^-_{\text{VDN}}(\mathbf{s}_{t+1}, \mathbf{a}') = Q^-_{\text{AKI}}(s_{t+1}^{\text{AKI}}, a'^{\text{AKI}}) + Q^-_{\text{SEP}}(s_{t+1}^{\text{SEP}}, a'^{\text{SEP}})$. The TD objective is $\mathcal{L}_{\text{TD}} = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}, \text{done}_t) \sim \mathcal{D}_{\text{com}}}[(Q_{\text{VDN}}(\mathbf{s}_t, \mathbf{a}_t) - y_t)^2]$. This additive decomposition preserves separate disease-conditioned value branches while allowing them to optimize the shared comorbid outcome. However, standard VDN propagates the same joint TD residual to both branches without specifying how much of the update should be assigned to each disease at a given patient state. BIFOCAL addresses this ambiguity through the dynamic risk–conflict credit-routing mechanism described in Appendix B.2.

B.2 Training Procedure

Algorithm 1 summarizes the learning pipeline. Stage I initializes the AKI and Sepsis heads from disease-specific Q-learning experts, giving each head a clinically meaningful warm start before joint optimization. Stage II trains the frozen routing modules: R_ϕ estimates disease-specific future deterioration from the historical window, while C_ψ estimates disagreement over the shared fluid decision from pretrained disease preferences. Stage III then performs coordinated value learning. For each minibatch, the frozen routing modules produce risk and conflict scores, their interaction gives disease-specific update weights, and the shared TD error is routed asymmetrically while the conflict-weighted anchor prevents unnecessary drift from the pretrained experts.

B.3 Decision Procedure

Algorithm 2 gives the decision-time procedure used to obtain an executable treatment. The frozen risk and conflict modules are queried from the current history and pretrained disease-reference outputs. When conflict is below c_0, the disease heads receive equal influence because their shared-fluid

Algorithm 1 BIFOCAL Training Procedure

Input: Aligned trajectories $\mathcal{D}$; discount factor γ; past and future windows K, H; risk temperature τ_r; anchor coefficient λ_{anchor}
Output: Disease-specific Q-functions $Q^{\text{AKI}}, Q^{\text{SEP}}$ and frozen routers R_ϕ, C_ψ

Stage I: Disease-Specific Pretraining
for $d \in \{\text{AKI}, \text{SEP}\}$ **do**
  Train a disease-specific Q-learning expert Q^d_{pre} on the disease-specific view $\mathcal{D}_d$
  Initialize $Q^d \leftarrow Q^d_{\text{pre}}$
end for
Stage II: Risk and Conflict Estimation
for each transition index t in $\mathcal{D}$ **do**
  Construct history $\mathcal{H}_t = (\mathbf{s}_{t-K+1}, \ldots, \mathbf{s}_t)$
  Construct future deterioration labels y^d_t from $(\mathbf{s}_{t+1}, \ldots, \mathbf{s}_{t+H})$, for each disease d
end for
Train R_ϕ to estimate $(\rho^{\text{AKI}}_t, \rho^{\text{SEP}}_t) = R_\phi(\mathcal{H}_t)$
Derive fluid-preference distributions $p^d_t(\cdot)$ from the frozen Q^d_{pre}
Train C_ψ to predict normalized fluid-preference disagreement $C_t \in [0, 1]$
Freeze R_ϕ and C_ψ
Stage III: Coordinated Policy Learning
for each minibatch $\mathcal{B} \subset \mathcal{D}$ **do**
  $(\rho^{\text{AKI}}_t, \rho^{\text{SEP}}_t) \leftarrow R_\phi(\mathcal{H}_t)$
  $C_t \leftarrow C_\psi(h^{\text{AKI}}_{t,\text{pre}}, h^{\text{SEP}}_{t,\text{pre}}, \hat{a}^{\text{AKI}}_{t,\text{pre}}, \hat{a}^{\text{SEP}}_{t,\text{pre}})$
  $\mathbf{w}_t \leftarrow \text{softmax}\big(C_t[\rho^{\text{AKI}}_t, \rho^{\text{SEP}}_t]/\tau_r\big)$
  $Q_{\text{tot}} \leftarrow Q^{\text{AKI}} + Q^{\text{SEP}}$
  $y_t \leftarrow r_t + \gamma(1 - \text{done}_t) \max_{\mathbf{a}'} Q^-_{\text{tot}}(\mathbf{s}_{t+1}, \mathbf{a}')$
  for $d \in \{\text{AKI}, \text{SEP}\}$ **do**
    $Q^d_t \leftarrow Q^d(s^d_t, a^d_t)$
    $Q^d_{t,\text{route}} \leftarrow \text{sg}(Q^d_t) + w^d_t[Q^d_t - \text{sg}(Q^d_t)]$
    $\mathcal{L}^d_{\text{anchor}} \leftarrow \text{KL}\big(\pi^d_{\text{pre}}(\cdot \mid s^d_t) \,\|\, \pi^d(\cdot \mid s^d_t)\big)$
  end for
  $\mathcal{L}_{\text{MARL}} \leftarrow \big(Q^{\text{AKI}}_{t,\text{route}} + Q^{\text{SEP}}_{t,\text{route}} - y_t\big)^2 + \lambda_{\text{anchor}}(1 - C_t) \sum_d \mathcal{L}^d_{\text{anchor}}$
  Update $Q^{\text{AKI}}, Q^{\text{SEP}}$ using $\nabla \mathcal{L}_{\text{MARL}}$
end for
return $Q^{\text{AKI}}, Q^{\text{SEP}}, R_\phi, C_\psi$

preferences are already aligned. When conflict is high, the decision weights follow the same risk–conflict softmax used during learning. BIFOCAL then fuses disease-specific Q-values only along the shared fluid dimension, selects the coordinated fluid level, and finally selects the disease-specific ultrafiltration and vasopressor actions conditioned on that fluid decision.

B.4 Future-Oriented Deterioration Targets

At each ICU decision time t, the model uses the past K patient states as input, from $\mathbf{s}_{t-K+1}$ to $\mathbf{s}_t$. The future window contains the next H states, from $\mathbf{s}_{t+1}$ to $\mathbf{s}_{t+H}$, and is used only during training to construct supervision targets. Rather than using static diagnostic labels, we define two guideline-anchored trajectory targets, y^{AKI}_t and y^{SEP}_t, to summarize future worsening along the renal and septic axes.

The AKI marker set is based on KDIGO-aligned renal dysfunction and critical-care AKI severity markers, including *serum creatinine*, *BUN*, *urine output*, *cumulative fluid balance*, and *RRT initiation* (Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group, 2012; Bouchard et al., 2009). The Sepsis marker set is based on Sepsis-3 and Surviving Sepsis Campaign severity dimensions, including *lactate*, *mean arterial pressure*, *vasopressor intensity*, *SOFA score*, and *shock index* (Singer et al., 2016; Evans et al., 2021; Middleton et al., 2019).

Algorithm 2 BIFOCAL Decision Procedure

Input: Current comorbid state $\mathbf{s}_t$; history $\mathcal{H}_t$; trained $Q^{\mathrm{AKI}}, Q^{\mathrm{SEP}}$; frozen R_ϕ, g_η; conflict threshold c_0; routing temperature τ_r.
Output: Executable treatment action $\mathbf{a}_t^*$

Compute $(\rho_t^{\mathrm{AKI}}, \rho_t^{\mathrm{SEP}}) \leftarrow R_\phi(\mathcal{H}_t)$
Compute pretrained representations $h_{t,\mathrm{pre}}^{\mathrm{AKI}}, h_{t,\mathrm{pre}}^{\mathrm{SEP}}$
Compute pretrained greedy proposals $\hat{a}_{t,\mathrm{pre}}^d = \arg\max_{a^d} Q_{\mathrm{pre}}^d(s_t^d, a^d)$
Compute $C_t \leftarrow g_\eta(h_{t,\mathrm{pre}}^{\mathrm{AKI}}, h_{t,\mathrm{pre}}^{\mathrm{SEP}}, \hat{a}_{t,\mathrm{pre}}^{\mathrm{AKI}}, \hat{a}_{t,\mathrm{pre}}^{\mathrm{SEP}})$
if $C_t < c_0$ **then**
    $\mathbf{w}_t^{\mathrm{dec}} \leftarrow (\frac{1}{2}, \frac{1}{2})$
else
    $\mathbf{w}_t^{\mathrm{dec}} \leftarrow \mathrm{softmax}\big(C_t[\rho_t^{\mathrm{AKI}}, \rho_t^{\mathrm{SEP}}]/\tau_r\big)$
end if
for each fluid level $a_{\mathrm{F}} \in \mathcal{A}_{\mathrm{F}}$ **do**
    $Q_t^{\mathrm{AKI,F}}(a_{\mathrm{F}}) \leftarrow \max_{a_{\mathrm{UF}}} Q^{\mathrm{AKI}}(s_t^{\mathrm{AKI}}, (a_{\mathrm{F}}, a_{\mathrm{UF}}))$
    $Q_t^{\mathrm{SEP,F}}(a_{\mathrm{F}}) \leftarrow \max_{a_{\mathrm{Vaso}}} Q^{\mathrm{SEP}}(s_t^{\mathrm{SEP}}, (a_{\mathrm{F}}, a_{\mathrm{Vaso}}))$
    $Q_{\mathrm{fuse}}^{\mathrm{F}}(\mathbf{s}_t, a_{\mathrm{F}}) \leftarrow w_t^{\mathrm{dec,AKI}} Q_t^{\mathrm{AKI,F}}(a_{\mathrm{F}}) + w_t^{\mathrm{dec,SEP}} Q_t^{\mathrm{SEP,F}}(a_{\mathrm{F}})$
end for
$a_{t,\mathrm{F}}^* \leftarrow \arg\max_{a_{\mathrm{F}}} Q_{\mathrm{fuse}}^{\mathrm{F}}(\mathbf{s}_t, a_{\mathrm{F}})$
$a_{t,\mathrm{UF}}^* \leftarrow \arg\max_{a_{\mathrm{UF}}} Q^{\mathrm{AKI}}(s_t^{\mathrm{AKI}}, (a_{t,\mathrm{F}}^*, a_{\mathrm{UF}}))$
$a_{t,\mathrm{Vaso}}^* \leftarrow \arg\max_{a_{\mathrm{Vaso}}} Q^{\mathrm{SEP}}(s_t^{\mathrm{SEP}}, (a_{t,\mathrm{F}}^*, a_{\mathrm{Vaso}}))$
$\mathbf{a}_t^* \leftarrow (a_{t,\mathrm{F}}^*, a_{t,\mathrm{UF}}^*, a_{t,\mathrm{Vaso}}^*)$
return $\mathbf{a}_t^*$

For each disease $d \in \{\mathrm{AKI}, \mathrm{SEP}\}$, we compute an averaged future deterioration score Δ_t^d by measuring marker changes from the current state to the future window, orienting all variables so that larger values indicate worsening, and normalizing variables with different units. The final target is $y_t^d = \sigma(\Delta_t^d / T)$, where T controls target smoothness. Because these targets are derived from future severity changes rather than static diagnostic labels, the risk module learns the trajectory direction of disease-specific deterioration.

B.5 Two-Stage Risk Module Training

To obtain stable temporal representations before optimizing disease-specific deterioration scores, we train the risk module in two stages. The first stage is label-free with respect to disease risk and learns a future-oriented representation from trajectory dynamics; the second stage uses this representation to predict AKI- and Sepsis-specific deterioration scores.

In the first stage, the risk heads are frozen, and only the temporal representation pathway is optimized. Given a historical trajectory $\mathcal{H}_t$, a past-window encoder E_p first produces $h_t^p = E_p(\mathcal{H}_t)$, which is mapped by a projection module P_p to a predicted future latent state $z_t^{\mathrm{pred}} = P_p(h_t^p)$. During training, the observed future window $\mathcal{F}_t$ provides a self-supervised target: a future-window encoder E_f and projector P_f produce $h_t^f = E_f(\mathcal{F}_t)$ and $z_t^{\mathrm{future}} = P_f(h_t^f)$. The prediction loss $\mathcal{L}_{\mathrm{pred}} = \mathrm{MSE}(z_t^{\mathrm{pred}}, \mathrm{sg}(z_t^{\mathrm{future}}))$ encourages the historical trajectory to encode information predictive of the upcoming patient course, where $\mathrm{sg}(\cdot)$ denotes stop-gradient. To ensure that z_t^{future} remains tied to the actual future trajectory rather than becoming an arbitrary latent target, a decoder D_f reconstructs the observed future window from this future latent, $\widehat{\mathcal{F}}_t = D_f(z_t^{\mathrm{future}})$, with $\mathcal{L}_{\mathrm{recon}} = \mathrm{MSE}(\widehat{\mathcal{F}}_t, \mathcal{F}_t)$. The first-stage objective is therefore $\mathcal{L}_{\mathrm{future}} = \mathcal{L}_{\mathrm{pred}} + \mathcal{L}_{\mathrm{recon}}$.

In the second stage, the learned future-oriented representation is aligned with disease-specific deterioration supervision. Two risk heads are trained on top of z_t^{pred}: the AKI head outputs $o_t^{\mathrm{AKI}} = G_{\mathrm{AKI}}(z_t^{\mathrm{pred}})$, and the Sepsis head outputs $o_t^{\mathrm{SEP}} = G_{\mathrm{SEP}}(z_t^{\mathrm{pred}})$. These logits are converted into continuous risk scores by sigmoid activation, $\rho_t^{\mathrm{AKI}} = \sigma(o_t^{\mathrm{AKI}})$ and $\rho_t^{\mathrm{SEP}} = \sigma(o_t^{\mathrm{SEP}})$. Since the supervision targets y_t^{AKI} and y_t^{SEP} are continuous future deterioration scores, the disease-

Dataset	Rows	Patients	Train	Test
MIMIC-IV	180,683	15,979	9,587	6,392
MIMIC-III	113,584	8,751	5,250	3,501
eICU	66,488	4,722	2,833	1,889
Private cohort	7,236	402	343	59

Table 6: Comorbid aligned cohort statistics. Rows denote aligned 4-hour AKI–Sepsis decision steps. Train and test columns report patient counts. MIMIC-III, MIMIC-IV, and eICU use patient-level 60/40 splits with fixed random state 0; the private cohort follows the package split, with original train and validation patients combined as training and original test patients held out.

specific losses are $\mathcal{L}_{\mathrm{AKI}} = \mathrm{SmoothL1}(\rho_t^{\mathrm{AKI}}, y_t^{\mathrm{AKI}})$ and $\mathcal{L}_{\mathrm{SEP}} = \mathrm{SmoothL1}(\rho_t^{\mathrm{SEP}}, y_t^{\mathrm{SEP}})$. The second-stage objective combines disease-specific risk prediction with the auxiliary future-modeling regularizer, $\mathcal{L}_{\mathrm{risk}} = \lambda_{\mathrm{future}}\mathcal{L}_{\mathrm{future}} + \lambda_{\mathrm{AKI}}\mathcal{L}_{\mathrm{AKI}} + \lambda_{\mathrm{SEP}}\mathcal{L}_{\mathrm{SEP}}$, where the λ terms balance representation preservation and risk-score prediction. After training, inference uses only the historical trajectory pathway, producing z_t^{pred} from $\mathcal{H}_t$ and then outputting ρ_t^{AKI} and ρ_t^{SEP}.

C Reproducibility Details

C.1 Randomness and Seed Control

All stochastic components were run with explicit seeds. Model training scripts expose a command-line `-seed` argument, which is used to initialize Python's `random` module, NumPy, and PyTorch through `random.seed`, `np.random.seed`, and `torch.manual_seed`. PyTorch deterministic algorithms are enabled where supported. Dataset splitting and bootstrap-based OPE evaluation also use fixed random states: data splits are generated with fixed `random_state` values, and reference OPE/bootstrap estimates use fixed state 42 for `-reference-random-state` or `-random-state`. All seeds required to reproduce the reported experiments are provided in the released codebase, including the experiment scripts and saved model metadata/checkpoints.

C.2 Computing Infrastructure

Experiments were conducted on a macOS workstation with an Apple M1 Max processor, 10 CPU cores, 64 GB unified memory, and an integrated 32-core Apple GPU. The operating system was macOS 14.5. The Python environment used Python 3.10.20 with PyTorch 2.12.0, NumPy 2.2.6, pandas 2.3.3, scikit-learn 1.7.2, SciPy 1.15.3, and Matplotlib 3.10.8. All reported experiments were executed on CPU unless explicitly stated otherwise.

C.3 Datasets, Preprocessing, Cohort Construction, and Train/Test Splits

Datasets We evaluate the framework on four ICU datasets. MIMIC-IV is used as the primary development and main evaluation cohort. MIMIC-III provides an additional public ICU cohort processed with the same 4-hour trajectory interface. eICU is used as a multicenter external validation cohort after conversion into the same MIMIC-like state schema. A private cohort is used as an additional external validation dataset. Across all datasets, patients are represented as aligned 4-hour trajectories and are converted into paired AKI and Sepsis views for the comorbid MARL task.

Comorbid cohort construction For MIMIC-style datasets, AKI and Sepsis disease-specific cohorts are first built separately and then intersected by ICU stay identifier. The comorbid cohort keeps shared mortality information and disease-specific onset summaries. If the AKI and Sepsis source cohorts contain different mortality indicators for the same ICU stay, the implementation uses the maximum mortality value, so that any recorded death is treated as death. The default comorbid onset time is the earliest of the AKI and Sepsis onset times. For the private cohort, the supplied package is already a sepsis-source cohort; therefore, the first source time step is used as the sepsis onset reference, and the supplied train/validation/test split is preserved.

Temporal alignment and views All trajectories use a 4-hour decision interval. AKI and Sepsis rows are aligned on $(\texttt{icustayid}, \texttt{timestep})$, and only rows present in both disease views are retained. Each retained row has two disease-conditioned observations, s_t^{AKI} and s_t^{SEP}, corresponding to the same patient and the same 4-hour decision step. The MIMIC-IV AKI view contains 101 variables and the Sepsis view contains 47 variables. Both views include demographics, vital signs, laboratory measurements, ventilation indicators, severity scores, fluid input/output, and cumulative fluid balance. The AKI view additionally retains renal and RRT-related variables such as creatinine, BUN, urine output, ultrafiltration, blood-flow and dialysis-related fields, and renal-support indicators. The Sepsis view emphasizes hemodynamic and sepsis-severity variables such as lactate, mean arterial pressure, vasopressor dose, SOFA, SIRS, shock index, and oxygenation.

Missing fields and normalization All datasets are mapped to the same feature schema. When an external dataset lacks a source variable required by the MIMIC-IV schema, the missing source column is filled with zero before the normalized model files are generated. Normalization is fitted on the training split only and is then reused for the test split. As-is variables are converted to numeric values and shifted by 0.5. Continuous variables are standardized with a training-set `StandardScaler`. Log-normalized variables are transformed as $\log(0.1 + \max(x, 0))$ before standardization. Remaining missing values after transformation are set to zero. The fitted scaler and the feature-column groups are saved with each disease view to keep train/test and cross-dataset preprocessing consistent.

Action construction Actions are constructed at the same 4-hour resolution. The shared fluid variable is `input_step`, representing total equivalent IV fluid within the interval. Continuous infusions are integrated by overlap with the 4-hour window, boluses are added if their timestamps fall inside the window, and all amounts are aggregated into interval-level volume. In the default comorbid build, the Sepsis `input_step` is used as the shared-fluid source and is copied into both disease views before action binning. The AKI-specific secondary action is `Ultrafiltrate_Output`; the Sepsis-specific secondary action is `max_dose_vaso`, the maximum norepinephrine-equivalent vasopressor dose in the interval. Negative action-source rows are removed before binning.

Each treatment dimension is discretized into five levels. The first level corresponds to no treatment, and the four positive levels are determined from nonzero training-set doses using empirical quartiles. A disease-specific action is encoded as $5a_F + a_{\mathrm{sec}}$, yielding 25 AKI actions and 25 Sepsis actions. The executable joint action is encoded as $25a_F + 5a_{UF} + a_{\mathrm{Vaso}}$, yielding 125 legal joint actions with one shared fluid level, one ultrafiltration level, and one vasopressor level.

C.4 Model and Baseline Implementations

Shared implementation settings All neural baselines use the same aligned train/test data, action discretization, terminal reward magnitude 100, discount factor $\gamma = 0.99$, five bins per treatment dimension, sparse-autoencoder state dimension 256, DQN hidden dimension 128, batch size 64, and soft target-network update rate 0.01, unless otherwise stated. The terminal reward is $+100$ for survival and -100 for death, with zero reward at nonterminal time steps. Model selection and reference metrics use the saved action-bin metadata from the training split.

Compared method groups The learned comparators are organized into two groups. (1) Multi-agent RL: Independent Q-Learning (Tan 1993), VDN (Sunehag et al. 2018), and QMIX (Rashid et al. 2018) are standard value-based MARL methods, while CFCQL (Shao et al. 2023) and B3C (Kim and Sycara 2026) are offline/conservative MARL methods; together these five test whether existing MARL methods can coordinate disease-specific agents without the specialized risk–conflict credit routing used by BIFOCAL. (2) Single-disease methods: HRRT (Xu et al. 2026) and WD3QNE (Wu et al. 2023) are adapted to the same five-bin treatment interface to test whether methods designed around one clinical decision process can transfer to comorbid AKI–Sepsis control without explicit disease-agent decomposition.

Clinician policy The clinician row is the behavior-policy reference estimated from observed clinician actions in the training split. For off-policy evaluation, the implementation fits an AI-Clinician-style reference model by clustering normalized states into 750 discrete states and estimating empirical clinician action probabilities. Rare transitions are filtered with transition threshold 5. The

clinician policy is evaluated on the same held-out trajectories and reward definition as learned policies.

IQL Independent Q-learning trains separate AKI and Sepsis dueling DQN agents. The AKI agent observes the AKI view and predicts fluid–ultrafiltration actions; the Sepsis agent observes the Sepsis view and predicts fluid–vasopressor actions. Each branch uses its own sparse autoencoder and prioritized replay buffer. At joint evaluation, the two independently selected disease actions are mapped into one executable 125-action treatment by combining the two disease-specific fluid recommendations through the default action-space mean rule and retaining the AKI ultrafiltration and Sepsis vasopressor components.

VDN VDN uses the same two disease-specific dueling DQN branches as IQL but trains them jointly with an additive value decomposition,

$$Q_{\text{joint}}(\mathbf{s}_t, \mathbf{a}_t) = Q_{\text{AKI}}(s_t^{\text{AKI}}, a_t^{\text{AKI}}) + Q_{\text{SEP}}(s_t^{\text{SEP}}, a_t^{\text{SEP}}).$$

The VDN baseline does not use the future-risk model, conflict router, adaptive credit routing, or conflict-gated Q fusion. Its decision rule is the uniform action-mean fusion used as the non-adaptive joint-action baseline.

QMIX QMIX replaces the additive VDN sum with a monotonic neural mixing network. The mixer receives the two scalar disease-agent Q-values and a global state formed by concatenating the AKI and Sepsis latent state vectors. Hypernetworks generate nonnegative mixing weights, preserving monotonicity of the joint value with respect to each agent value. The default mixing hidden dimension is 32. The QMIX baseline does not use the proposed risk–conflict routing components.

CFCQL CFCQL is implemented as a conservative VDN-style baseline. It keeps the two-agent VDN backbone and adds a conservative counterfactual value regularizer. For each minibatch sample, the implementation enumerates one agent's actions while holding the other agent's dataset action fixed, computes a temperature-scaled log-sum-exp counterfactual value, and subtracts the dataset joint Q-value. The two one-agent counterfactual terms are averaged and weighted by $\alpha = 1.0$ with temperature 1.0.

B3C B3C is implemented on the QMIX backbone. The implementation includes the two B3C ingredients compatible with this discrete clinical setting: critic target clipping and behavior-cloning regularization. The TD target is clipped symmetrically to the terminal reward range, and a cross-entropy behavior-cloning loss is applied to the AKI and Sepsis per-agent Q logits using the observed clinician actions. The default behavior-cloning weight is 1.0 and the behavior-cloning temperature is 1.0.

HRRT HRRT is implemented as a single-disease baseline over the repository's 5×5 disease-specific action space. Each disease view trains its own sparse autoencoder and a 25-action BCQ/CQL network with separate Q and imitation heads. The imitation head defines the BCQ support mask by retaining actions whose relative behavior probability exceeds 0.3, while the Q head is trained with a clipped TD target, cross-entropy imitation loss, and conservative log-sum-exp penalty. The default conservative weight is 0.1, the CQL temperature is 1.0, the imitation and Q losses both have weight 1.0, and rare positive secondary actions are oversampled toward a target ratio of 0.08. For the comorbid evaluation, the AKI HRRT model predicts fluid–ultrafiltration actions, the Sepsis-view HRRT model predicts fluid–vasopressor actions, and the pair is mapped into the same executable 125-action joint treatment used by the MARL baselines.

WD3QNE WD3QNE is a direct joint dueling DQN baseline. Instead of training separate disease agents, it concatenates the AKI and Sepsis latent states, standardizes the combined representation, and clusters it into a default of 120 discrete state IDs using MiniBatch k-means. A dueling Q network with a 128-dimensional state embedding predicts all 125 joint actions directly. During target construction, the next action is chosen from the observed clinician transition when the current SOFA score is below 4.0 and from the greedy learned action otherwise; the resulting target is clipped to $[-100, 100]$. The default training uses 8 epochs, batch size 256, learning rate 10^{-4}, target-update rate 0.20, and a smoothed empirical behavior table with $\alpha = 10^{-3}$. At evaluation time, target probabilities are produced by a softmax over the 125 joint-action Q-values with temperature 1.0.

Our model BIFOCAL uses the VDN-style disease-specific value branches but adds the proposed risk–conflict learning and decision mechanisms. The model is initialized from an IQL pair when warm-starting is enabled. A future-risk model predicts AKI and Sepsis deterioration risks from the past four 4-hour states over a future six-step, 24-hour horizon. A frozen conflict router estimates current AKI–Sepsis action conflict from frozen disease-agent Q profiles, using Jensen–Shannon divergence between the two fluid-marginal action distributions as supervision. During joint training, the shared TD update is routed by conflict-gated disease-risk weights, and an IQL-anchor KL loss is weighted by $1 - C_t$. At decision time, the default rule is `risk_gated_q`: risk-weighted Q fusion is activated only when conflict exceeds the threshold, otherwise the fluid decision remains balanced.

C.5 Off-Policy Evaluation

Behavior and target probabilities All OPE metrics compare a learned target policy π with the estimated clinician behavior policy μ. For single-agent evaluation, $\mu(a_t \mid s_t)$ is read from the AI-Clinician reference model for the corresponding disease view. For joint evaluation, AKI and Sepsis behavior probabilities are combined into a 125-action joint behavior distribution using the same executable-action mapping as the target policy. Learned policies are converted to target probabilities by applying row-wise softmax to Q-values, or to the fused 125-action Q-values for Q-level fusion.

WIS and WISt For trajectory i, the trajectory-level importance ratio is defined as $\rho_i = \prod_{t=0}^{T_i-1} \frac{\pi(a_{i,t}|s_{i,t})}{\mu(a_{i,t}|s_{i,t})}$, and the discounted return is $G_i = \sum_t \gamma^t r_{i,t}$. The implementation computes ρ_i in log space for numerical stability. Weighted importance sampling is computed as $\widehat{V}_{\mathrm{WIS}} = \frac{\sum_i \rho_i G_i}{\sum_i \rho_i}$. WISt uses the same estimator after truncating each raw trajectory weight at 100, namely $\widehat{V}_{\mathrm{WISt}} = \frac{\sum_i \min(\rho_i,100) G_i}{\sum_i \min(\rho_i,100)}$.

Bootstrap WIS metrics and effective sample size WISb and WISbt are bootstrap estimates of WIS and WISt. The default evaluation uses patient-level bootstrap resamples and random state 42; confidence intervals are the 2.5 and 97.5 percentiles of the bootstrap distribution. The effective sample size for untruncated weights is computed from normalized trajectory weights $\bar{\rho}_i = \rho_i / \sum_j \rho_j$ as $\mathrm{ESS} = 1/\sum_i \bar{\rho}_i^2$. For truncated weights $w_i^t = \min(\rho_i, 100)$, the truncated effective sample size is $\mathrm{ESSt} = (\sum_i w_i^t)^2 / \sum_i (w_i^t)^2$.

Doubly robust estimate The doubly robust estimate combines fitted Q-values with importance-weighted temporal-difference corrections. Let $V_\pi(s_t) = \sum_a \pi(a \mid s_t) Q(s_t, a)$. For one trajectory, the implemented estimate is $\widehat{V}_{\mathrm{DR}} = V_\pi(s_0) + \sum_{t=0}^{T-1} \gamma^t \left(\prod_{k=0}^{t} \rho_k \right) [r_t + \gamma V_\pi(s_{t+1}) - Q(s_t, a_t)]$. Per-step ratios are clipped to $[0.1, 10]$, and the resulting trajectory estimate is clipped to the terminal reward range when reward clipping is enabled. The reported DR value is the mean over patient trajectories.

Treatment discrepancy and action consistency $\mathrm{RMSE}_{\mathrm{IV}}$, $\mathrm{RMSE}_{\mathrm{UF}}$, and $\mathrm{RMSE}_{\mathrm{Vaso}}$ measure disagreement between clinician and learned treatment intensity after decoding discrete actions back to their training-set median dose values. In joint evaluation, the fluid component is decoded from the 125-action joint action, ultrafiltration from the AKI secondary component, and vasopressor dose from the Sepsis secondary component. PF1 is patient-level exact-action agreement: exact-match accuracy is computed within each patient trajectory and then averaged across patients. SF1 is sample-level exact-action agreement over all aligned test rows. Thus PF1 weights patients equally, whereas SF1 weights decision steps equally.

C.6 Clinician Evaluation and Ethics

Results As shown in **Fig. 13**, 6 ICU physicians (two each at junior, intermediate, and senior levels) evaluated 30 comorbid cases, including 10 AKI-dominant, 10 Sepsis-dominant, and 10 balanced cases. Before viewing model outputs, each physician independently rated AKI versus Sepsis severity, providing a clinical reference for BIFOCAL's disease priorities. Physicians were then shown ρ_t^{AKI} and ρ_t^{SEP} (predicted future deterioration risks), C_t (current fluid-preference conflict), $\mathbf{w}_t$ (re-



Figure 13: **Clinician evaluation.** Six ICU clinicians rated BIFOCAL across 30 cases on a 7-point scale. S1–S10, A1–A10, and B1–B10 denote individual Sepsis-dominant, AKI-dominant, and balanced comorbid cases, respectively. **a,** Component-disease priority. **b–f,** Interpretability, expected effectiveness, safety awareness, trust, and adoption willingness.

sulting disease-priority weights), and the coordinated treatment, and rated **disease-priority appropriateness**, **interpretability**, **expected effectiveness**, **safety awareness**, **trust**, and **adoption willingness**. Disease-priority appropriateness and interpretability received the highest ratings (5.5–6.3), while other dimensions were also rated positively. Ratings were similar across physician experience levels. BIFOCAL's priorities were consistent with physicians' assessments made before seeing the model, and positive post-reveal ratings further support the clinical plausibility of its disease prioritization and coordinated treatment.

Study Design and Scope We conducted a retrospective physician evaluation to assess whether BIFOCAL's disease-priority signals and coordinated treatment recommendations were clinically plausible, interpretable, and potentially useful as decision-support information. This simulated evaluation did not affect actual patient care and was not designed to establish prospective effectiveness, causal treatment benefit, or clinical safety.

Physician Participants The study included six ICU physicians from four hospitals. Physician experience was classified as junior (1–3 years of clinical experience), intermediate (4–7 years), senior (8–19 years), or highly experienced (a department director or at least 20 years of clinical experience). For consistency with the three experience groups reported in the main paper, senior and highly experienced physicians were combined into the senior group. The six-physician panel therefore comprised two junior, two intermediate, and two senior physicians.

None of the participating physicians was involved in model development, model training, case selection, or result analysis. All physicians evaluated the cases independently and were not permitted to discuss individual cases during the study. Hospital names and other identifying information are withheld during anonymous review.

Clinical-Stratum Definition and Case Sampling The physician study used a stratified random sample of 30 decision points from the held-out MIMIC-III test set, comprising 10 Sepsis-dominant, 10 AKI-dominant, and 10 balanced cases. The selected patients and decision points were not used for model training, hyperparameter tuning, or checkpoint selection. For each eligible patient, one decision point was randomly selected to avoid repeated observations from the same patient.

Before case sampling, eligible decision points were classified by independent physicians who were not involved in model development or the subsequent recommendation-rating study. They reviewed only the clinical information available up to each decision point and rated the relative immediate treatment priority on a five-level scale: -2 (clearly Sepsis-prioritized), -1 (moderately Sepsis-

prioritized), 0 (approximately balanced), $+1$ (moderately AKI-prioritized), and $+2$ (clearly AKI-prioritized).

The median rating was used to define Sepsis-dominant (≤ -1), balanced ($= 0$), and AKI-dominant ($\geq +1$) strata. Cases with unresolved disagreement were assessed by an additional independent adjudicator according to the same criteria.

BIFOCAL outputs, recorded clinician actions after the decision point, and subsequent patient outcomes were unavailable during stratification. Ten cases were then randomly sampled from each stratum using a fixed random seed. Thus, the three strata represent independently assessed clinical priorities rather than priorities inferred from BIFOCAL.

Two-Stage Evaluation Procedure Each of the six evaluating physicians independently reviewed all 30 selected cases. Case order was randomized separately for each physician.

In the first stage, physicians reviewed only the clinical history available up to the evaluated decision point. They did not see BIFOCAL outputs, the stratum label, the recorded clinician action, or subsequent patient outcomes. Each physician independently rated the relative immediate priority of AKI and Sepsis using the same five-level scale. These pre-reveal ratings were retained for evaluating agreement between physicians' independent clinical assessments and BIFOCAL's disease-priority weights; they did not change the previously assigned case strata.

In the second stage, physicians were shown BIFOCAL's predicted AKI deterioration risk ρ_t^{AKI}, predicted Sepsis deterioration risk ρ_t^{SEP}, present fluid-preference conflict C_t, resulting disease-priority weights $(w_t^{\mathrm{AKI}}, w_t^{\mathrm{SEP}})$, and final coordinated treatment recommendation. They then rated the model output across six dimensions. Each physician completed the evaluation independently, producing $6 \times 30 = 180$ physician–case evaluations.

Information Presented to Physicians For each case, physicians received the same standardized summary of the patient's clinical history up to the decision point, including the available physiological measurements and previous treatments. Future clinical information, subsequent outcomes, and the treatment recorded after the decision point were hidden to avoid hindsight and imitation bias.

During the second stage, the two predicted deterioration risks were described as model-estimated risks of near-future AKI and Sepsis deterioration. The conflict score was described as the model-estimated disagreement between the two disease-specific fluid preferences, and the priority weights indicated the relative influence assigned to the two disease-oriented branches. The recommended fluid, ultrafiltration, and vasopressor levels were then displayed as one executable joint treatment.

Rating Instrument After viewing the model outputs, physicians rated the following six statements on a seven-point Likert scale:

1. **Disease-priority appropriateness**: The displayed priority appropriately reflects which disease should receive greater treatment emphasis at this decision point.
2. **Interpretability**: The displayed risk, conflict, and priority information makes the treatment recommendation understandable.
3. **Expected effectiveness**: The recommended treatment is clinically likely to support the intended treatment goals.
4. **Safety awareness**: The recommendation appropriately reflects foreseeable risks and potential cross-disease treatment conflicts.
5. **Trust**: I would trust this recommendation as supporting information for clinical decision making.
6. **Adoption willingness**: I would be willing to consider this recommendation when making a treatment decision.

The scale ranged from 1 (strongly disagree or completely inappropriate) to 7 (strongly agree or completely appropriate), with 4 indicating a neutral or uncertain judgment. Expected-effectiveness and safety-awareness scores represent physicians' retrospective judgments rather than observed treatment effects or prospective safety evidence.

Ethics and Limitations The evaluation used retrospective, de-identified MIMIC-III records and did not affect clinical care. Participating physicians provided consent to take part in the recommendation-rating study. The study followed the applicable institutional ethics-review procedures, with identifying approval information withheld during anonymous review.

This evaluation has limitations. It included six physicians and 30 retrospective cases and therefore cannot establish prospective effectiveness or safety. The results should consequently be interpreted as evidence of perceived clinical plausibility, interpretability, and potential decision-support value rather than clinical benefit.