AGENTIFYING PATIENT DYNAMICS WITHIN LLMS THROUGH INTERACTING WITH CLINICAL WORLD MODEL: A CASE STUDY OF SEPSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sepsis management in the ICU requires sequential treatment decisions under rapidly evolving patient physiology. Although large language models (LLMs) encode broad clinical knowledge and can reason over guidelines, they are not inherently grounded in action-conditioned patient dynamics. We introduce **SepsisAgent**, a world model-augmented LLM agent for sepsis treatment recommendation. SepsisAgent uses a learned Clinical World Model to simulate patient responses under candidate fluid–vasopressor interventions, and follows a propose–simulate–refine workflow before committing to a prescription. We first show that world-model access alone yields inconsistent LLM decision performance, motivating agent-specific training. We then train SepsisAgent through a three-stage curriculum: patient-dynamics supervised fine-tuning, propose–simulate–refine behavior cloning, and world-model-based agentic reinforcement learning. On MIMIC-IV sepsis trajectories, SepsisAgent outperforms all traditional RL and LLM-based baselines in off-policy value while achieving the best safety profile under guideline adherence and unsafe-action metrics. Further analysis shows that repeated interaction with the Clinical World Model enables the agent to learn regularities in patient evolution, which remain useful even when simulator access is removed.

1 INTRODUCTION

Sepsis remains a leading cause of mortality in Intensive Care Units (ICUs), presenting a major challenge in critical care medicine Singer et al. (2016); Rudd et al. (2020); Rhee et al. (2017). Effective management requires clinicians to titrate intravenous fluids and vasopressors over time to restore perfusion while avoiding downstream organ injury. This makes sepsis treatment a high-stakes *sequential decision-making* problem: actions that improve short-term hemodynamics may still worsen long-term outcomes, for example when aggressive fluid resuscitation restores blood pressure but increases the risk of pulmonary edema or renal failure Dobson et al. (2024); Meyhoff et al. (2022); Douglas et al. (2020). Clinicians must make these decisions under substantial cognitive load, integrating high-dimensional physiological streams Helman et al. (2022) while accounting for patient-specific heterogeneity Komorowski et al. (2018); Seymour et al. (2019). This motivates decision-support systems that combine clinical knowledge with patient dynamics Boussina et al. (2024).

Large Language Models (LLMs) offer a promising interface for clinical decision support because they can interpret heterogeneous clinical context, reason over medical guidelines, and provide natural-language rationales Xu et al. (2025b); Sellergren et al. (2025); Chen et al. (2024). However, LLMs are not inherently grounded in patient dynamics. They may know that vasopressors increase blood pressure in general, but still fail to estimate how a specific patient will respond to a specific dose over the next few hours Wornow et al. (2023). This limitation is especially problematic in sepsis, where the quality of a treatment decision depends not only on whether it is guideline-consistent, but also on how it changes the patient's future trajectory.

A natural way to address this limitation is to augment the LLM with a predictive model of patient dynamics. We define a **Clinical World Model** as a learned, action-conditioned approximation of

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

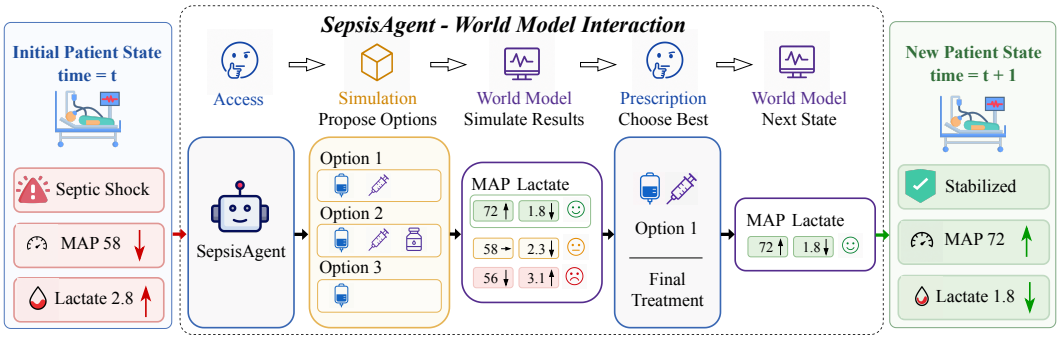


Figure 1: Illustration of SepsisAgent's propose–simulate–refine workflow. Given the current patient state, the agent proposes candidate fluid–vasopressor actions, queries the World Model for predicted patient responses, and commits to a final treatment action based on the simulated trajectories.

patient evolution. Given a patient state and a candidate treatment action, the world model predicts possible physiological responses and downstream outcome signals. This gives the LLM a mechanism to compare counterfactual treatment options before committing to a prescription. Yet world-model access alone is insufficient: predicted trajectories are approximate, and a generic LLM may over-trust short-term simulated improvements or misinterpret noisy feedback. Therefore, the key challenge is not only to build a simulator, but to train an agent that can use simulated patient responses correctly.

To this end, we introduce **SepsisAgent**, a world model-augmented LLM agent for sepsis treatment recommendation. SepsisAgent follows a propose–simulate–refine workflow: it proposes candidate fluid–vasopressor interventions, queries the Clinical World Model for simulated patient responses, and refines its final prescription using both simulated feedback and clinical priors. We further train the agent through a three-stage curriculum. First, supervised fine-tuning teaches patient-dynamics prediction and guideline-aware one-step reasoning. Second, behavior cloning teaches multi-round interaction with the world model. Third, world-model-based agentic reinforcement learning optimizes long-horizon treatment strategies through simulated state–action–outcome feedback.

Our contributions are summarized as follows:

- **A world model-augmented LLM agent for sepsis treatment.** We propose **SepsisAgent**, an LLM-based treatment agent that collaborates with a Clinical World Model to compare candidate fluid–vasopressor interventions before prescription. SepsisAgent outperforms all traditional RL and LLM-based baselines on MIMIC-IV sepsis trajectories.
- **A world-model-based multi-stage training paradigm.** We show that merely giving LLMs access to a world model is insufficient. We therefore train SepsisAgent through a multi-stage agentic learning pipeline based on the Clinical World Model, improving its ability to interpret simulated patient responses and optimize treatment policies.
- **Internalizing patient dynamics within LLMs.** Our experiments show that world-model-based agentic reinforcement learning enhances the LLM's intrinsic ability to predict patient dynamics in sepsis, rather than merely fitting the reward signal. SepsisAgent improves prediction of in-hospital mortality and vasopressor requirement even without simulator access.

2 Towards World Model Augmented Sepsis Agent

2.1 Problem Definition

AI agents are increasingly used to support healthcare decision-making. In this work, we focus on sepsis treatment, a challenging ICU task characterized by rapidly worsening organ dysfunction caused by infection Singer et al. (2016). Following prior studies that formulate sepsis treatment as a discrete Markov Decision Process (MDP) Komorowski et al. (2018); Raghu et al. (2017), we model this task as a sequential decision-making problem. At each decision step t, the agent observes the current patient state $s_t \in \mathcal{S}$, represented by demographic, physiological, laboratory, and treatment-history variables, and selects a treatment action $a_t \in \mathcal{A}$.

Following prior sepsis RL formulations Wu et al. (2023); Kalimouttou et al. (2025), our action space covers two key controllable hemodynamic interventions: intravenous fluid administration and vasopressor use, and discretizes these interventions by dosage levels into a 5×5 grid:

$$\mathcal{A} = \mathcal{A}^{\text{fluid}} \times \mathcal{A}^{\text{vaso}}.$$

The agent aims to recommend treatment actions that stabilize short-term patient physiology while improving long-term clinical outcomes.

2.2 Motivation to Introduce Clinical World Model

Most prior methods learn treatment policies directly from retrospective EHR trajectories, mapping the current patient state to a treatment action Liu et al. (2020). Such policies mainly learn associations between observed states and historical clinician actions. They are therefore vulnerable to historical behavior bias and have limited ability to reason about alternative treatment paths that are poorly covered in observational data Xu et al. (2025a). However, sepsis treatment is not only about selecting an action for the current state; it also requires estimating how the patient may evolve under different interventions Raghu et al. (2018).

This motivates an explicit model of patient response. In model-based reinforcement learning, a world model approximates environment dynamics and supports planning by predicting future states under candidate actions Ha & Schmidhuber (2018); LeCun (2022). In this work, **we define a Clinical World Model as a learned, action-conditioned approximation of patient dynamics**:

$$W_\theta : (s_t, a_t) \mapsto p_\theta(s_{t+1}, o_t \mid s_t, a_t),$$

where s_t is the current patient state representation, a_t is a candidate treatment action, and o_t denotes downstream clinical outcomes. For sepsis treatment, this model acts as an approximate simulator that predicts the physiological consequences of fluid–vasopressor interventions, enabling the agent to compare candidate actions before making a prescription.

2.3 Solution: World Model-Augmented Sepsis Agent

LLMs and Clinical World Models address different parts of the treatment-decision problem. LLMs provide the semantic layer: they encode clinical prior knowledge, interpret heterogeneous patient context, reason over guidelines, and generate interpretable rationales for treatment decisions Singhal et al. (2023; 2025); Tu et al. (2025). However, they do not directly model how a specific patient state will evolve after a specific intervention Steinberg et al. (2024); Wornow et al. (2024). Clinical World Models provide the dynamics layer: they predict action-conditioned patient responses under candidate treatments, but lack the clinical semantic reasoning and guideline-aware judgment of LLMs.

To combine these strengths, we propose **SepsisAgent**, a world model-augmented LLM agent for sepsis treatment recommendation. The LLM backbone acts as the decision-making policy, while the Clinical World Model acts as an agent component that collaborates with the LLM by predicting patient responses under candidate interventions. Moreover, the world model provides counterfactual exploration trajectories, which are used to further train SepsisAgent to understand patient dynamics and optimize treatment strategies.

At inference time, SepsisAgent follows a propose–simulate–refine workflow. Given the current patient state s_t, the LLM proposes a candidate action set

$$\mathcal{C}_t = \{a_t^{(1)}, \ldots, a_t^{(M)}\}.$$

For each candidate action, the Clinical World Model estimates its possible physiological consequence:

$$\hat{y}_t^{(i)} \sim W_\theta(s_t, a_t^{(i)}),$$

where $\hat{y}_t^{(i)}$ summarizes the predicted patient response, including future physiological state and outcome-related signals. The LLM then compares these action–response pairs with its clinical priors and commits to the final treatment:

$$a_t = \pi_\phi\left(s_t, \{(a_t^{(i)}, \hat{y}_t^{(i)})\}_{i=1}^{M}\right).$$

Building on this formulation, **Section 3 instantiates the Clinical World Model and evaluates its benefit for SepsisAgent**. **Section 4 then describes how SepsisAgent explores and learns within this world model**.

3 Training World Model to Augment Sepsis Agent

This section builds the Clinical World Model that serves as both the inference-time simulator and the training environment for SepsisAgent. We first construct discrete sepsis trajectories from MIMIC-IV, then instantiate an action-conditioned predictive model of patient dynamics, and finally **evaluate whether this model provides useful feedback for treatment decision-making.**

We extract 20,092 ICU stays from MIMIC-IV Johnson et al. (2023) using Sepsis-3 criteria Singer et al. (2016), covering the window from 24h before to 48h after sepsis onset within each ICU stay. Following AI Clinician Komorowski et al. (2018), data are aggregated into 4-hour steps and split into training, validation, and test sets (7:2:1). The state space consists of 42 clinical variables (Table 6), and the action space is a 5×5 discrete grid based on dosage percentiles of intravenous fluids and vasopressors. Further details are provided in Appendix C.

3.1 Clinical World Model Instantiation

Following MedDreamer Xu et al. (2025a), we use a GRU encoder to encode and update latent representations of sepsis patient trajectories (shown in Figure 2). Specifically, given the observed trajectory history

$$\tau_{\leq t} = \{(s_1, a_1), \ldots, (s_{t-1}, a_{t-1}), s_t\},$$

a two-layer GRU encoder produces a history-aware patient representation:

$$h_t = \mathrm{GRU}_\psi(\tau_{\leq t}).$$

We adopt this lightweight architecture for proof-of-concept simplicity; as discussed in Appendix D.4, increasing model complexity or varying architecture size did not yield significant performance gains.

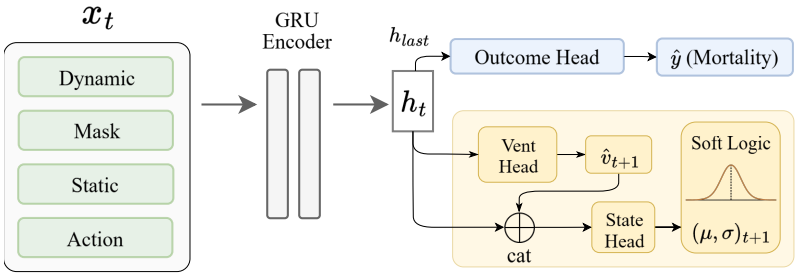


Figure 2: Architecture of the Clinical World Model.

The World Model is instantiated with two prediction heads. First, the state transition head predicts the next physiological state conditioned on the encoded patient history and candidate treatment action:

$$p_\theta(s_{t+1} \mid \tau_{\leq t}, a_t) = \mathcal{N}\left(\mu_\theta(h_t, a_t), \Sigma_\theta(h_t, a_t)\right).$$

We use the predicted mean $\hat{s}_{t+1} = \mu_\theta(h_t, a_t)$ as the simulated next-state response returned to the agent. To improve clinical plausibility, auxiliary heads predict ventilation status and derived severity scores, including SOFA and SIRS, which are used as regularization targets during training.

Second, the outcome head estimates longer-term clinical consequences over a fixed prediction window of length K. Given a trajectory segment

$$\tau_{t:t+K} = \{(s_t, a_t), \ldots, (s_{t+K}, a_{t+K})\},$$

the outcome model predicts the corresponding clinical outcome:

$$\hat{o}_{t:t+K} = \mathcal{W}_\phi^{\mathrm{outcome}}(\tau_{t:t+K}).$$

Together, the transition and outcome heads allow the world model to provide both short-horizon physiological responses and longer-horizon risk signals for candidate treatment actions. Detailed training configurations are provided in Appendix D.

3.2 Clinical World Model as a Predictive Simulator

We evaluate the Clinical World Model as a predictive simulator and report compact metrics characterizing state prediction accuracy and outcome discrimination (Table 1). The results show that the model captures clinical relevant patient dynamics and provides informative short-horizon feedback for candidate treatment actions.

Table 1: Clinical World Model Results.

Model Component	Metric	Value
State Transition	MAE	0.316
	Ventilation AUC	0.942
Outcome Prediction	AUC-ROC	0.804
	AUC-PR	0.663

3.3 Why World-Model Access Alone Is Insufficient

To quantify the benefit of the Clinical World Model, we evaluate three sota LLMs (GPT-4.1-mini, Gemini-3-Flash, and o3) under three settings: (1) **Vanilla LLM**; (2) **LLM + World Model**, where the model can query simulated patient responses for candidate actions; and (3) **LLM + World Model + Clinical Prior**, where the model additionally receives concise clinical priors derived from sepsis guidelines Evans et al. (2021). For efficiency and reproducibility, we randomly sample 725 episodes from the test set as the evaluation benchmark for LLM-based methods[1] (shown in Table 2).

Table 2: World-model access alone yields inconsistent LLM decision performance.

Method	Off-Policy Eval			Sepsis Guideline	Unsafe Actions (%)	
	DR (↑)	WIS (↑)	WPDIS (↑)	Adherence (% ↑)	Underdosing (↓)	Overdosing (↓)
Human Reference						
Clinicians (Test Set)	5.06	5.27	10.82	94.76	0.35	0.19
GPT-4.1-mini	6.13	6.59	10.82	80.59	0.66	2.18
w/ World Model	7.61$_+$	6.69$_+$	9.03$_-$	84.05$_+$	0.59$_+$	2.87$_-$
w/ World Model + Clinical Prior	7.31$_+$	5.21$_-$	17.09$_+$	94.00$_+$	0.58$_+$	1.60$_+$
Gemini-3-Flash	8.17	9.09	13.98	96.43	1.19	2.58
w/ World Model	4.05$_-$	10.01$_+$	11.33$_-$	93.62$_-$	1.90$_-$	2.60$_-$
w/ World Model + Clinical Prior	4.49$_-$	7.78$_-$	12.24$_-$	95.16$_-$	0.78$_+$	1.16$_+$
o3	8.32	9.17	20.38	90.55	0.72	1.57
w/ World Model	8.78$_+$	9.98$_+$	20.17$_-$	92.08$_+$	0.64$_+$	1.72$_-$
w/ World Model + Clinical Prior	9.46$_+$	10.27$_+$	22.95$_+$	96.91$_+$	0.09$_+$	0.24$_+$

Note: Detailed definitions of policy value, guideline adherence, and safety metrics are provided in Section 5.1. Green ($_+$) indicates improvement over Vanilla, while Red ($_-$) indicates degradation.

The results show that world model augmentation alone is insufficient: generic LLMs may misinterpret simulated patient responses and choose actions that improve short-term signals but hurt longer-term outcomes. Appendix I provides expert-validated failure-mode analysis for such cases.

Adding clinical priors from sepsis guidelines Evans et al. (2021) improves safety metrics in most cases, but does not guarantee policy-value improvement. For example, Gemini-3-Flash with world model and clinical prior performs worse than its vanilla version across all OPE metrics.

Key Finding. World-model access is not enough; the agent must learn how to use it.

This observation shifts our focus from world-model access to training an agent to use its feedback. In Section 4, we use simulated patient responses as learning signals to train SepsisAgent to understand patient dynamics, reason under sepsis guideline priors, and refine treatment policies through repeated interaction with the Clinical World Model.

4 Agentifying Patient Dynamics in Agents through World Model Interaction

4.1 From World Model Feedback to Patient-Dynamics Understanding

Training goal. Section 3.3 shows that the bottleneck is not simulator access, but the agent's ability to interpret patient-dynamics feedback and integrate it with clinical priors. From a clinical decision-making perspective, a useful agent should not simply follow the action with the best one-step sim-

[1]For Gemini-3-Flash with world-model augmentation, each episode rollout costs approximately $0.31.

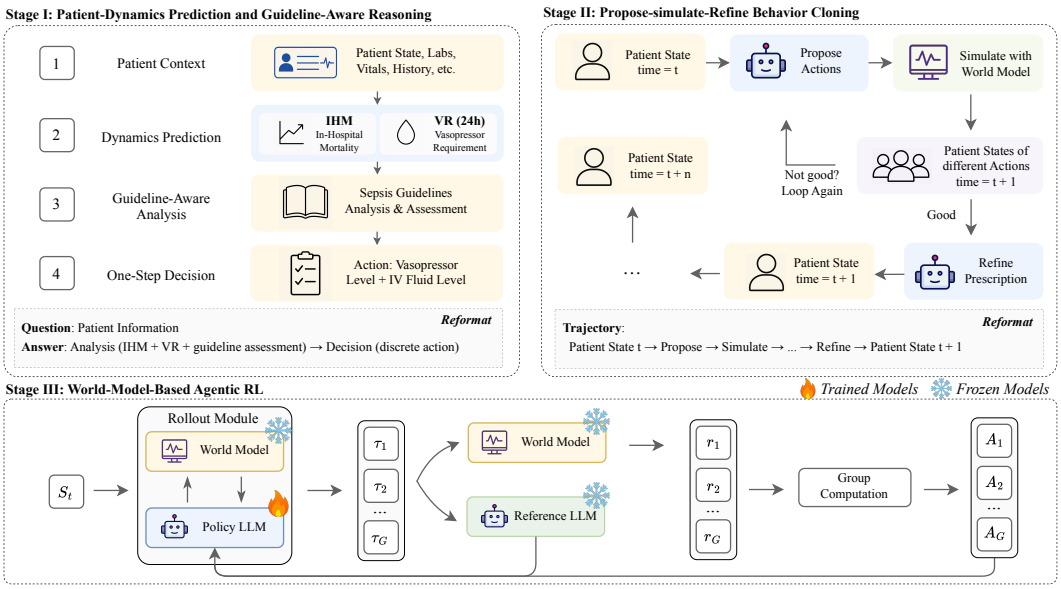


Figure 3: Overview of the three-stage training pipeline for SepsisAgent. ***Stage I*** focuses on guideline-aware patient state understanding, ***Stage II*** learns agentic interaction with a clinical world model, and ***Stage III*** applies GRPO, treating the world model as an environment to iteratively predict future patient states and refine treatment policies through rollout interactions.

ulated response. It should first understand the patient's evolving risk, then reason about guideline-consistent treatment choices, and finally refine decisions through counterfactual patient-response estimates. We therefore use the Clinical World Model not only as an inference-time simulator, but also as a training environment that exposes the LLM to repeated state–action–outcome feedback.

Curriculum design. This motivates a staged curriculum, shown in Figure 3. Stage I builds patient-dynamics understanding and guideline-aware treatment reasoning through supervised data that jointly includes IHM prediction, VR prediction, and one-step treatment recommendation. Stage II teaches multi-turn interaction with the world model through imitation learning on synthesized interaction trajectories. Stage III further optimizes the agent through RL in the world-model environment, encouraging long-horizon planning rather than greedy short-term stabilization.

4.2 Stage I: Patient-Dynamics Prediction and Guideline-Aware Reasoning

Dynamics-aware supervision. We construct supervised training data around two prediction tasks defined in the MIMIC-Sepsis benchmark Huang et al. (2025): in-hospital mortality (IHM) and vasopressor requirement (VR). These tasks are critical for treatment selection, as they capture long-term patient outcome and hemodynamic deterioration.

In-hospital mortality. Given the current patient context and treatment information, the model predicts the patient's subsequent in-hospital mortality as a binary outcome. This task encourages the model to attend to the long-term consequences of treatment trajectories rather than only immediate physiological measurements.

Vasopressor requirement. The VR task predicts whether the patient will require vasopressor support within the next 24 hours. Based on sepsis guidelines Evans et al. (2021), VR serves as a clinically meaningful proxy for predicting whether key physiological indicators will cross high-risk clinical thresholds within the next 24 hours.

One-step treatment reasoning. We train the model to perform single-step treatment reasoning grounded in patient dynamics and sepsis guideline priors. Each example follows an **Analysis–Decision** format: in the **Analysis** phase, the model predicts IHM and VR and evaluates the patient's condition under sepsis guideline priors; in the **Decision** phase, it recommends the corresponding

discrete treatment action. This design makes the two dynamics-prediction tasks part of the treatment reasoning process rather than isolated auxiliary objectives.

Reasoning data construction. Following common practice in constructing medical reasoning supervision Chen et al. (2024); Sun et al. (2025), we use GPT-4.1 to synthesize chain-of-thought-style rationales from observed clinical facts, expert actions, and sepsis guideline priors. The generated traces are reformatted into structured supervision pairs, enabling the model to learn patient-state analysis and guideline-consistent one-step treatment recommendation.

4.3 Stage II: Propose–Simulate–Refine Behavior Cloning

Workflow imitation. Stage I equips the model with non-agentic patient-dynamics prediction and guideline-aware one-step reasoning. Stage II teaches it to use world-model feedback during decision-making: before each treatment action, the agent can query the Clinical World Model over multiple rounds, inspect simulated consequences, and refine its prescription.

Trajectory synthesis. For each expert transition (s_t, a_t^*), we use GPT-4.1 to synthesize a multi-round propose–simulate–refine reasoning trace following the interaction in Section 2.3. The trace starts from the patient state s_t, proposes candidate actions $\mathcal{C}_t$, obtains simulated responses $\{\hat{y}_t^{(i)}\}_{i=1}^{M}$ from the Clinical World Model, optionally updates the candidate set based on these responses, and finally selects the expert action a_t^*.

Behavior cloning objective. We train the model to reproduce these structured multi-round traces. This provides a cold start for world-model interaction, enabling the agent to learn how simulated patient responses should inform treatment recommendations before Stage III reinforcement learning.

4.4 Stage III: World-Model-Based Agentic RL

Long-horizon optimization. Behavior cloning provides a cold start for propose–simulate–refine decision-making, but remains constrained by expert demonstrations. In Stage III, we further optimize the policy with Group Relative Policy Optimization (GRPO) Shao et al. (2024). The agent interacts with the Clinical World Model as a virtual environment, samples multiple reasoning trajectories for each patient state, and learns to select actions that improve long-term patient outcomes rather than merely imitate observed decisions.

World-model rollouts. For each initial patient state s_0, the policy π_ϕ generates treatment actions through multi-round world-model interaction. The Clinical World Model then rolls out the resulting patient trajectory

$$\tau = \{(s_0, a_0), (s_1, a_1), \ldots, (s_T, a_T)\},$$

where each transition is induced by the simulated patient response under the selected treatment action. This allows the agent to receive feedback from complete treatment trajectories, rather than only single-step expert labels.

Composite reward. We optimize the policy using a composite reward that combines terminal outcome, intermediate physiological stabilization, and guideline consistency:

$$R(\tau) = R_{\text{out}}(s_T) + \sum_{t=0}^{T-1} r(s_t, s_{t+1}) - \lambda_g \mathcal{P}_g(\tau).$$

Here, R_{out} rewards favorable terminal outcomes such as survival and penalizes mortality, while $\mathcal{P}_g$ penalizes violations of sepsis guideline constraints.

Intermediate reward. Following the intermediate reward design in DDQN for sepsis treatment Raghu et al. (2017), we define the step-wise physiological reward as

$$r(s_t, s_{t+1}) = C_0 \mathbb{I}(\Delta\text{SOFA}_t = 0 \wedge s_{t+1}^{\text{SOFA}} > 0) + C_1 \Delta\text{SOFA}_t + C_2 \Delta\text{Lac}_t.$$

This term encourages improvement in organ dysfunction and lactate dynamics, while the terminal reward encourages long-term survival. Detailed reward coefficients, lactate transformation, clipping strategy, and implementation details are provided in Appendix F.2.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Backbone selection. We select Qwen3-4B-Instruct Yang et al. (2025) as the backbone of SepsisAgent. It is sufficiently capable of supporting clinical reasoning and agentic interaction, while remaining efficient for deployment-style evaluation: after training, the average inference time per decision step is 6.1s, well within the minutes-scale decision workflow of ICU sepsis management.

Baselines. We compare SepsisAgent with four groups of baselines. **Clinicians** correspond to the recorded human decisions in the MIMIC-IV test set and serve as the real-world reference policy. **Traditional RL** baselines include representative sepsis treatment policies, DDQN Raghu et al. (2017) and AI Clinician Komorowski et al. (2018). **Vanilla LLMs** evaluate whether general-purpose language models can recommend sepsis treatments directly from patient states without world-model feedback. **World-model-augmented LLMs** use the same Clinical World Model and propose–simulate–refine interaction protocol as SepsisAgent, but do not receive our staged supervised and reinforcement learning training. This comparison isolates the effect of agent training from merely providing world-model access.

Evaluation metrics. We evaluate each method on the 725-episode held-out test set from three complementary perspectives. **Off-policy evaluation** reports DR Jiang & Li (2016), WIS, and WPDIS Precup et al. (2000) to estimate policy value from retrospective trajectories without real-world deployment. **Sepsis Guideline Adherence** measures whether recommended actions satisfy the sepsis guideline constraints used throughout this work Evans et al. (2021). **Unsafe Actions** further reports the percentage of extreme underdosing and overdosing actions, following rule-based safety definitions derived from expert clinical practice Festor et al. (2022). This metric is stricter than guideline adherence, used only as an independent evaluation criterion, and never optimized during training. Detailed rules are provided in Appendix E.

Table 3: SepsisAgent achieves the best overall value and safety profile.

Method	Off-Policy Eval			Sepsis Guideline	Unsafe Actions (%)	
	DR (↑)	WIS (↑)	WPDIS (↑)	Adherence (% ↑)	Underdosing (↓)	Overdosing (↓)
Human Reference						
Clinicians (Test Set)	5.06	5.27	10.82	94.76	0.35	<u>0.19</u>
Traditional RL						
DDQN Raghu et al. (2017)	8.69	6.19	15.11	82.79	0.67	1.01
AI Clinician Komorowski et al. (2018)	5.98	9.94	9.88	87.24	0.53	2.34
WD3QNE Wu et al. (2023)	8.72	**12.07**	<u>23.20</u>	87.60	1.11	1.49
Vanilla LLMs						
o3	8.32	9.17	20.38	90.55	0.72	1.57
Gemini-3-Pro Google DeepMind (2026)	5.84	8.59	19.68	96.74	<u>0.09</u>	1.62
Gemini-3-Flash Google DeepMind (2025)	8.17	9.09	13.98	96.43	1.19	2.58
GPT-OSS-120B OpenAI (2025b)	8.25	6.00	21.17	79.42	1.06	1.11
GPT4.1-mini OpenAI (2025a)	6.13	6.59	10.82	80.59	0.66	2.18
Deepseek-V3.2 DeepSeek-AI et al. (2025)	8.63	10.25	15.07	81.80	0.19	0.41
World-Model-Augmented LLMs						
o3 + WM	<u>9.46</u>	10.27	22.95	<u>96.91</u>	0.09	0.24
Gemini-3-Flash + WM	4.49	7.78	12.24	95.16	0.78	1.16
GPT4.1-mini + WM	7.32	5.21	17.09	94.00	0.58	1.60
SepsisAgent and Its Backbone						
Qwen3-4B-Instruct Yang et al. (2025)	7.79	7.34	18.76	78.00	0.62	2.13
SepsisAgent	**10.01**	<u>11.14</u>	**23.40**	**97.95**	**0.08**	**0.14**

5.2 MAIN RESULTS: SEPSISAGENT IMPROVES POLICY VALUE WHILE PRESERVING SAFETY

Table 3 summarizes the performance of SepsisAgent and all baselines on the 725-episode test set.

SepsisAgent achieves the strongest off-policy value. SepsisAgent obtains the best DR and WPDIS scores among all methods, while remaining competitive on WIS. It outperforms both traditional RL and LLM-based policies, and also improves consistently over its base model across

all three OPE estimators. This suggests that our staged training pipeline benefits from the LLM's clinical prior knowledge while further improving treatment policy quality.

SepsisAgent achieves the best safety profile. SepsisAgent achieves the highest sepsis guideline adherence and the lowest unsafe-action rates, indicating that the agent does not improve estimated policy value by taking unsafe treatment shortcuts. Traditional RL policies tend to produce more extreme overdosing actions. Clinician behavior is more conservative, but can still contain suboptimal underdosing patterns, consistent with prior observations in sepsis treatment practice Wu et al. (2023); Festor et al. (2022).

Training matters beyond world-model access. World-model-augmented LLMs improve over many vanilla LLM baselines, but they still fall short of SepsisAgent. This suggests that simulated patient responses alone are not enough. The agent must be trained to interpret dynamics feedback and align it with clinical safety constraints.

5.3 Ablation Study: RL Drives Policy Improvement

Table 4 reports the cumulative effect of each training stage. The OPE results show that policy-value improvement mainly comes from the RL stage. SFT and behavior cloning improve some metrics, but do not yield stable gains across DR, WIS, and WPDIS. The safety metrics improve earlier: Stage I and Stage II already teach the model to reason within sepsis guideline and expert-defined safety boundaries. Stage III further improves both safety metrics, suggesting that the Clinical World Model provides an effective RL training environment for optimizing long-horizon treatment value while preserving clinically safe behavior.

Table 4: Staged training improves policy value, safety, and internalization of patient dynamics.

Method	Off-Policy Eval (↑)			Sepsis Guideline	Unsafe	IHM (↑)		VR (↑)	
	DR	WIS	WPDIS	Adherence (% ↑)	Actions (% ↓)	AUROC	AUPRC	AUROC	AUPRC
Base Model									
Qwen3-4B-Instruct	7.79	<u>7.34</u>	18.76	78.00	2.75	65.27	45.01	70.62	61.74
SepsisAgent Variants									
SepsisAgent (Stage I: SFT)	<u>9.21</u>	7.17	19.56	88.01	1.09	67.50	<u>50.25</u>	<u>76.40</u>	<u>65.11</u>
SepsisAgent (Stage I+II: +BC)	8.99	6.81	<u>19.61</u>	<u>96.89</u>	<u>0.51</u>	**67.55**	46.63	74.56	63.70
SepsisAgent (Stage I+II+III: +RL)	**10.01**	**11.14**	**23.40**	**97.95**	**0.22**	<u>68.52</u>	**53.45**	**79.96**	**68.83**

5.4 Analysis: Internalizing Patient Dynamics within LLMs

We test whether SepsisAgent internalizes patient dynamics by evaluating IHM and VR prediction without access to the Clinical World Model. As shown in Table 4, Stage III achieves the best IHM AUPRC and the best VR AUROC/AUPRC, while maintaining competitive IHM AUROC. This suggests that world-model-based agentic RL improves the LLM's intrinsic ability to predict patient outcomes and future vasopressor needs. The agent therefore does not merely fit the reward signal; repeated interaction with the Clinical World Model helps it learn patient-dynamics regularities that remain useful even when simulator access is removed.

6 Conclusion

We presented **SepsisAgent**, a world model-augmented LLM agent for sepsis treatment recommendation. SepsisAgent uses a Clinical World Model to compare counterfactual treatment responses and refine fluid–vasopressor prescriptions. Our results show that world-model access alone is insufficient for generic LLMs, motivating a staged training pipeline that teaches the agent to interpret simulated patient dynamics and align them with clinical priors. On MIMIC-IV sepsis trajectories, SepsisAgent achieves the strongest off-policy value and the best safety profile among traditional RL and LLM-based baselines. Ablation studies further show that world-model-based RL drives policy-value improvement while preserving safety. Finally, intrinsic prediction results indicate that repeated interaction with the Clinical World Model helps the LLM internalize patient dynamics, improving mortality and vasopressor-requirement prediction even without simulator access. These findings provide a proof of concept for using world-model-based agentic learning to support sequential clinical decision-making in high-acuity settings.

AI USE STATEMENT

We used generative AI in this research. As described in the paper, GPT-4.1 was used to generate reasoning supervision for Stage I and multi-round interaction traces for Stage II. SepsisAgent and general-purpose language models were also evaluated as sepsis-treatment policies. In preparing the manuscript, AI tools assisted in drafting and editing portions of the text and tables. The authors reviewed the AI-assisted research outputs and manuscript and take responsibility for the final content of this paper.

REFERENCES

Aaron Boussina, Supreeth P Shashikumar, Atul Malhotra, Robert L Owens, Robert El-Kareh, Christopher A Longhurst, Kimberly Quintero, Allison Donahue, Theodore C Chan, Shamim Nemati, et al. Impact of a deep learning sepsis prediction model on quality of care and survival. *NPJ digital medicine*, 7(1):14, 2024.

Samuel M Brown, Michael J Lanspa, Jason P Jones, Kathryn G Kuttler, Yao Li, Rick Carlson, Russell R Miller III, Eliotte L Hirshberg, Colin K Grissom, and Alan H Morris. Survival after shock requiring high-dose vasopressor therapy. *Chest*, 143(3):664–671, 2013.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma, Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao,

Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL https://arxiv.org/abs/2512.02556.

Geoffrey P Dobson, Hayley L Letson, and Jodie L Morris. Revolution in sepsis: a symptoms-based to a systems-based approach? *Journal of Biomedical Science*, 31(1):57, 2024.

Ivor S Douglas, Philip M Alapat, Keith A Corl, Matthew C Exline, Lui G Forni, Andre L Holder, David A Kaufman, Akram Khan, Mitchell M Levy, Gregory S Martin, et al. Fluid response evaluation in sepsis hypotension and shock: a randomized clinical trial. *Chest*, 158(4):1431–1445, 2020.

Elham Estiri and Hossein Mirinejad. Model-free reinforcement learning for automated fluid administration in critical care. *arXiv preprint arXiv:2401.06299*, 2024.

Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flávia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Critical care medicine*, 49(11):e1063–e1143, 2021.

Paul Festor, Yan Jia, Anthony C Gordon, A Aldo Faisal, Ibrahim Habli, and Matthieu Komorowski. Assuring the safety of ai-based clinical decision support systems: a case study of the ai clinician for sepsis treatment. *BMJ health & care informatics*, 29(1):e100549, 2022.

Google DeepMind. Gemini 3 Flash model card, 2025. URL https://deepmind.google/models/model-cards/gemini-3-flash/.

Google DeepMind. Gemini 3 Pro model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-pro/.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Stephanie Helman, Martha Ann Terry, Tiffany Pellathy, Andrew Williams, Artur Dubrawski, Gilles Clermont, Michael R Pinsky, Salah Al-Zaiti, and Marilyn Hravnak. Engaging clinicians early during the development of a graphical user display of an intelligent alerting system at the bedside. *International journal of medical informatics*, 159:104643, 2022.

Yong Huang, Zhongqi Yang, and Amir Rahmani. Mimic-sepsis: A curated benchmark for modeling and learning from sepsis trajectories in the icu. In *2025 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 1–7. IEEE, 2025.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pp. 652–661. PMLR, 2016.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Alexandre Kalimouttou, Jason N Kennedy, Jean Feng, Harvineet Singh, Suchi Saria, Derek C Angus, Christopher W Seymour, and Romain Pirracchio. Optimal vasopressin initiation in septic shock: the oviss reinforcement learning study. *Jama*, 333(19):1688–1698, 2025.

Taylor W Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden parameter markov decision processes. *Advances in neural information processing systems*, 30, 2017.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Yann LeCun. A path towards autonomous machine intelligence. https://openreview.net/pdf?id=BZ5a1r-kVsf, 2022. Version 0.9.2, June 27, 2022.

Siqi Liu, Kay Choong See, Kee Yuan Ngiam, Leo Anthony Celi, Xingzhi Sun, and Mengling Feng. Reinforcement learning for clinical decision support in critical care: comprehensive review. *Journal of medical Internet research*, 22(7):e18477, 2020.

Siqi Liu, Qianyi Xu, Zhuoyang Xu, Zhuo Liu, Xingzhi Sun, Guotong Xie, Mengling Feng, and Kay Choong See. Reinforcement learning to optimize ventilator settings for patients on invasive mechanical ventilation: retrospective study. *Journal of Medical Internet Research*, 26:e44494, 2024.

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, 2024.

Tine S Meyhoff, Peter B Hjortrup, Jørn Wetterslev, Praleene Sivapalan, Jon H Laake, Maria Cronhjort, Stephan M Jakob, Maurizio Cecconi, Marek Nalos, Marlies Ostermann, et al. Restriction of intravenous fluid in icu patients with septic shock. *New England Journal of Medicine*, 386(26): 2459–2470, 2022.

Peter C Nauka, Jason N Kennedy, Emily B Brant, Matthieu Komorowski, Romain Pirracchio, Derek C Angus, and Christopher W Seymour. Challenges with reinforcement learning model transportability for sepsis treatment in emergency care. *npj Digital Medicine*, 8(1):91, 2025.

OpenAI. Gpt-4.1. `https://openai.com/index/gpt-4-1/`, 2025a.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025b. URL `https://arxiv.org/abs/2508.10925`.

Dilruk Perera, Siqi Liu, Kay Choong See, and Mengling Feng. Smart imitator: Learning from imperfect clinical decisions. *Journal of the American Medical Informatics Association*, 33(1): 49–66, 2026.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766, 2000.

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

Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602*, 2018.

Chanu Rhee, Raymund Dantes, Lauren Epstein, David J Murphy, Christopher W Seymour, Theodore J Iwashyna, Sameer S Kadri, Derek C Angus, Robert L Danner, Anthony E Fiore, et al. Incidence and trends of sepsis in us hospitals using clinical vs claims data, 2009-2014. *Jama*, 318(13):1241–1249, 2017.

Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, 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.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Christopher W Seymour, Jason N Kennedy, Shu Wang, Chung-Chou H Chang, Corrine F Elliott, Zhongying Xu, Scott Berry, Gilles Clermont, Gregory Cooper, Hernando Gomez, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *Jama*, 321(20):2003–2017, 2019.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. Motor: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Deli Zhao, Wenbing Huang, Tingyang Xu, Qifeng Bai, et al. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26446–26467, 2025.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025.

Jason Waechter, Anand Kumar, Stephen E Lapinsky, John Marshall, Peter Dodek, Yaseen Arabi, Joseph E Parrillo, R Phillip Dellinger, Allan Garland, Cooperative Antimicrobial Therapy of Septic Shock Database Research Group, et al. Interaction between fluids and vasoactive agents on mortality in septic shock: a multicenter, observational study. *Critical care medicine*, 42(10): 2158–2168, 2014.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023.

Michael Wornow, Suhana Bedi, Miguel Angel Fuentes Hernandez, Ethan Steinberg, Jason Alan Fries, Christopher Ré, Sanmi Koyejo, and Nigam H Shah. Context clues: Evaluating long context models for clinical prediction tasks on ehrs. *arXiv preprint arXiv:2412.16178*, 2024.

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.

Qianyi Xu, Gousia Habib, Feng Wu, Dilruk Perera, and Mengling Feng. Meddreamer: Model-based reinforcement learning with latent imagination on complex ehrs for clinical decision support. *arXiv preprint arXiv:2505.19785*, 2025a.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A Related Work

A.1 LLM Agent Using a World Model

Sepsis treatment involves complex and patient-specific hemodynamic dynamics, and current practices in high-acuity clinical settings are known to be suboptimal Nauka et al. (2025). The heterogeneous and temporally evolving nature of patient responses makes it difficult to determine optimal interventions using static clinical guidelines alone Steinberg et al. (2024); Huang et al. (2025).

To address this challenge, we draw inspiration from the notion of world models in model-based reinforcement learning Ha & Schmidhuber (2018); LeCun (2022) and instantiate it as a predictive simulator of patient physiology. In our framework, an LLM serves as the decision-making agent and queries the Clinical World Model during action selection. Given candidate treatment strategies, the world model provides approximate simulations of short-term physiological responses in sepsis patients. This predictive information complements the LLM's clinical reasoning and supports more informed treatment decisions.

A.2 Model-based Agentic Reinforcement Learning

Model-based reinforcement learning has been widely explored for treatment recommendation Xu et al. (2025a); Raghu et al. (2018); Perera et al. (2026), as it enables policy optimization through explicit modeling of patient dynamics and simulated rollouts. Compared to model-free approaches Liu et al. (2024); Estiri & Mirinejad (2024); Luo et al. (2024), model-based methods offer improved sample efficiency and the ability to reason about the consequences of treatment decisions in settings where real-world interaction is not feasible.

Building on this paradigm, we further construct the Clinical World Model as a virtual environment for training the LLM policy. Within this environment, the LLM performs simulated rollouts of treatment decisions and observes the resulting physiological trajectories Xu et al. (2025a); Raghu et al. (2018); Killian et al. (2017).

Through repeated interaction with the world model, the LLM can move beyond purely supervised imitation by incorporating simulated feedback into its decision process. Rather than treating model predictions as an oracle, this formulation allows the policy to reason over predicted trajectories in conjunction with its clinical priors. Moreover, the model-based setup naturally supports decision-making over multiple time steps, facilitating the consideration of both short-term physiological responses and longer-term clinical objectives.

B Limitations

This study is conducted solely for academic research purposes. The simulation results and models presented in this paper are not intended for real-world clinical decision-making. Any potential clinical use should be conducted only under the supervision of doctors.

C Data Sources Details

We use data from the Medical Information Mart for Intensive Care IV (MIMIC-IV, v2.2) Johnson et al. (2023). We identify eligible patients using the Sepsis-3 criteria Singer et al. (2016), selecting ICU stays with suspected infection (antibiotics and blood cultures within ± 24 hours) and a SOFA score ≥ 2. We exclude patients under 18, stays less than 24 hours, and cases with missing mortality data.

Data Preprocessing. We align our data extraction and preprocessing pipeline with the AI Clinician study Komorowski et al. (2018). Given the schema differences between MIMIC-III (used in prior work) and MIMIC-IV, we resolved variable mapping conflicts by referencing the official `mimic-code` repository[2]. Unlike previous works that use irregular time steps or 1-hour intervals,

[2] `https://github.com/MIT-LCP/mimic-code`

we aggregate data into 4-hour time steps to align with the clinical decision-making rhythm for stable sepsis management.

- **Feature Aggregation Logic:** Data is aggregated into 4-hour time steps starting from ICU admission ($t = 0$).
  - *Summation:* Total Effective Volume (TEV) and Urine Output are summed within the window. Cross-window fluid infusions are allocated proportionally based on duration.
  - *Maximum:* Vasopressors (NE-Eq) use the maximum infusion rate within the window. Ventilation status takes the most severe state (Invasive $>$ NIV $>$ HFNC $>$ O2 $>$ None).
  - *Mean/Worst:* Vital signs (HR, MAP, SpO_2) are averaged, while GCS uses the worst (minimum) value to capture neurological decline.
  - *Last Value:* Laboratory results use the last recorded value (sample-and-hold). SOFA scores are recalculated post-aggregation.
- **Imputation Strategy:** Missing values are handled via forward filling. If no prior value exists (e.g., at the start of an admission), we impute using the population median derived from the training set.

Action Space Definition. We define the two action dimensions as follows (detailed in Table 7):

1. **Norepinephrine Equivalent (NE-Eq):**Brown et al. (2013) Aggregates multiple vasopressors into a standard norepinephrine scale. Note that Dopamine is excluded from our final action space due to its declining usage in modern sepsis protocols.
2. **Total Effective Volume (TEV):**Waechter et al. (2014) Aggregates crystalloids and colloids based on their volume expansion effect. Dextrose 5% is excluded as it functions as free water rather than a volume expander.

Table 5: Cohort statistics stratified by 90-day mortality.

Group	% Female	Mean Age	Avg Steps	Population
Survivors	42.8	61.9	11.7	13,446
Non-survivors	43.0	68.0	11.5	6,646

Table 6: State observation variables ($D = 42$) grouped by physiological system. Data is aggregated into 4-hour time steps.

Category	Variables
Demographics (5)	Age, Gender, Weight, ICU Readmission Status, Elixhauser Comorbidity Index
Vital Signs (8)	Heart Rate, Mean Arterial Pressure (MAP), Respiratory Rate, SpO_2, Temperature, Glasgow Coma Scale (GCS) Total, Shock Index, Urine Output
Metabolic & Renal (10)	pH, Lactate, Bicarbonate, Base Excess, Anion Gap, BUN, Creatinine, Sodium, Potassium, Chloride, Glucose
Hematology (7)	Hemoglobin, Hematocrit, White Blood Cell (WBC) Count, Platelet Count, INR, PT, PTT
Organ Function (6)	Total Bilirubin, Albumin, ALT, AST, PaO_2/FiO_2 Ratio, SOFA Score (Total & Sub-scores)
Respiratory (3)	Mechanical Ventilation Status, FiO_2, $PaCO_2$
Others (3)	Total CO_2, Calcium (Total/Ionized), Magnesium

Table 7: Action space aggregation and discretization. The agent selects a discrete level (0-4) for both dimensions simultaneously.

Action Dimension	Unit	Aggregation Formula
1. Vasopressors (Norepinephrine Equivalent)	mcg/kg/min (NE-Eq)	$\text{NE-Eq} = \text{Norepinephrine} + \text{Epinephrine} + \text{Phenylephrine}/10 + \text{Dopamine}/100 + \text{Vasopressin} \times 2.5/60$
2. IV Fluids (Total Effective Volume)	mL/4h (TEV)	$\text{TEV} = \sum_k w_k V_k$ (See Appendix Table 8 for full ItemID curation) $+2 \times (V_{\text{Albumin 5\%}} + 5 \times V_{\text{Albumin 25\%}})$

Table 8: Detailed curation of Total Effective Volume (TEV) coefficients. We mapped over 20 distinct MIMIC-IV ItemIDs to their physiological expansion coefficients (w_k), ranging from 0.25 (hypotonic) to 8.0 (hypertonic).

Coeff (w_k)	**Fluid Type**	**MIMIC-IV ItemIDs**
$0.25\times$	Saline 0.255%	220958
$0.30\times$	Saline 0.3%	220959
$0.50\times$	NaCl 0.45%, D5 1/2NS	225159, 225823, 220965
$1.00\times$	NaCl 0.9%, LR, Plasma-Lyte	225158, 225828, 226372, etc.
$2.00\times$	Albumin 5%, FFP, Platelets	220864, 220970, 225168, 225170, 225171, 221000, 221013
$2.75\times$	Mannitol	227531
$3.00\times$	NaCl 3%	225161
$5.00\times$	Albumin 25%	220862
$6.66\times$	Sodium Bicarbonate 8.4%	220995, 227533
$8.00\times$	NaCl 23.4%	228341

D World Model Training Details

D.1 Architecture Details

We implement the Clinical World Model using PyTorch. The framework consists of a shared temporal encoder that extracts joint representations for both state transition modeling and outcome prediction.

- **Shared Backbone:** The backbone is a 2-layer Gated Recurrent Unit (GRU) with a hidden dimension of 128. To prevent overfitting, a dropout rate of 0.2 is applied between the GRU layers. The encoder takes a concatenated input of dynamic features, missingness masks, embedded static features (32-dimensional), and embedded treatment actions (32-dimensional).
- **State Prediction Head:** This head follows a hierarchical structure. It first predicts the ventilation status $\hat{v}_{t+1}$ via a Multi-Layer Perceptron (MLP). The hidden state h_t, augmented with the predicted ventilation status, is then projected to the parameters of a Gaussian distribution (μ, σ) for each of the dynamic physiological variables.
- **Outcome Head:** The outcome head is an MLP with a hidden layer of size 64 and ReLU activation. It takes the joint temporal representation h_{last} from a 48-hour trajectory window to predict the probability of 90-day mortality.
- **Soft Logic Layer:** To maintain clinical consistency, a differentiable Soft Logic layer reverses the normalization and log-transformations of the predicted states. It computes SOFA and SIRS scores using sigmoid-based soft thresholds with a temperature parameter $\tau = 10.0$.

D.2 Training and Optimization

The model is trained end-to-end by minimizing a composite multi-task loss function:

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \lambda_1 \mathcal{L}_{\text{Outcome}} + \lambda_2 \mathcal{L}_{\text{Reg}} + \lambda_3 \mathcal{L}_{\text{Vent}} \tag{1}$$

where $\mathcal{L}_{\text{NLL}}$ is the Gaussian negative log-likelihood of the next state, $\mathcal{L}_{\text{Outcome}}$ is the binary cross-entropy loss for mortality, and $\mathcal{L}_{\text{Reg}}$ is the Smooth-L1 loss for SOFA/SIRS consistency.

We use the **AdamW** optimizer with a learning rate of $1e-3$ and a weight decay of $1e-4$. The learning rate is dynamically adjusted using a `ReduceLROnPlateau` scheduler with a reduction factor of 0.5 and a patience of 3 epochs. Training is conducted with a batch size of 2048 for 50 epochs, utilizing early stopping with a patience of 8 epochs based on validation performance.

D.3 Hyperparameter Settings

For the shared clinical world model, we used a two-layer GRU backbone with a hidden dimension of 128 and a dropout rate of 0.2. Static patient features and treatment actions were separately projected into 32-dimensional embeddings before being integrated into the temporal modeling framework. The trajectory window size was set to $K = 12$, corresponding to a 48-hour observation window. During training, we used a batch size of 2048 and optimized the model with AdamW using a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. The loss weights were set to $\lambda_1 = 1.0$ for outcome prediction, $\lambda_2 = 0.01$ for consistency regularization, and $\lambda_3 = 0.3$ for ventilation prediction. Early stopping was applied with a patience of 8 epochs to prevent overfitting.

D.4 Performance of Different World Models

We compare four candidate architectures for the Clinical World Model: Transformer, LSTM, 2-layer GRU, and 8-layer GRU. Table 9 reports predictive performance of the world models themselves, while Table 10 reports the off-policy evaluation results of SepsisAgent trained with each corresponding world model.

Table 9: Predictive performance of different Clinical World Model architectures.

World Model	**State MAE** (↓)	**Ventilation AUC** (↑)	**Outcome AUC-ROC** (↑)	**Outcome AUC-PR** (↑)
Transformer	**0.292**	**0.951**	**0.821**	**0.686**
LSTM	<u>0.299</u>	0.940	0.805	<u>0.670</u>
GRU-8layers	0.308	<u>0.949</u>	<u>0.810</u>	0.657
GRU-2layers	0.316	0.942	0.804	0.663

Table 10: Off-policy evaluation of SepsisAgent trained with different Clinical World Models.

Method	**Off-Policy Eval** (↑)		
	DR	**WIS**	**WPDIS**
SepsisAgent-Transformer WM	**12.06**	**12.10**	22.10
SepsisAgent-LSTM WM	9.87	<u>10.62</u>	**23.50**
SepsisAgent-GRU-8layers WM	10.13	<u>11.99</u>	22.95
SepsisAgent-GRU-2layers WM	10.01	11.14	<u>23.40</u>

Tables 9 and 10 show that the Transformer world model achieves the best predictive performance across all metrics, but SepsisAgent trained with it does not show a consistent OPE advantage over recurrent world models. Nevertheless, we observe that more accurate world models tend to bring better DR and WIS improvements, suggesting that future Clinical World Models with higher predictive fidelity may further benefit SepsisAgent. Since the 2-layer GRU world model is already sufficient for SepsisAgent to outperform all baselines in the main experiments, we adopt this simplified architecture for proof-of-concept simplicity.

E Details of Sepsis Guidelines and Safety Metrics

Sepsis guideline priors. We use the following sepsis guideline priors derived from the *Surviving Sepsis Campaign 2021* guidelines Evans et al. (2021) and clinical expert review:

- Sepsis and septic shock are medical emergencies, and treatment and resuscitation should begin immediately.
- For patients with sepsis-induced hypoperfusion or septic shock, at least 30 mL/kg of IV crystalloid fluid should be given within the first 3 hours of resuscitation.
- For adults with septic shock on vasopressors, an initial target Mean Arterial Pressure (MAP) of 65 mmHg is recommended over higher targets.
- Vasopressors should be initiated if MAP remains below 65 mmHg after adequate fluid resuscitation.

Unsafe action metrics. In addition to guideline adherence, we report expert-defined unsafe-action rates following rule-based safety evaluation protocols Festor et al. (2022). These rules are stricter than the guideline-adherence metric and are used only for evaluation, never as training objectives. We define two unsafe-action types:

- **Extreme underdosing:** the patient is severely hypotensive, but the recommended treatment provides no vasopressor support and no or low IV fluid. Formally, this occurs when

$$\mathrm{MAP} < 55\ \mathrm{mmHg}, \quad a^{\mathrm{vaso}} = 0, \quad a^{\mathrm{fluid}} \leq \mathrm{Low}.$$

- **Extreme overdosing:** the patient is hypertensive, but the recommended treatment still gives a high vasopressor dose. Formally, this occurs when

$$\mathrm{MAP} > 95\ \mathrm{mmHg}, \quad a^{\mathrm{vaso}} > \mathrm{High}.$$

F Details of the Three-stage Training pipeline

This section provide a detailed description of the training data construction for *Stage I* and *Stage II*, followed by the GRPO-based agentic reinforcement learning procedure.

F.1 Data Construction

F.1.1 *Stage I* Training Data

ICU data contain a large amount of numerical state information. To help the LLM better understand and analyze such data in an aligned manner, we introduce this training stage and construct targeted training data accordingly. Specifically, each data instance consists of structured ICU clinical signals $\mathcal{S}$, including multivariate physiological and laboratory measurements over a fixed time window, together with a task-specific instruction. The training data are constructed by prompting GPT-4.1 to perform diverse instruction-driven tasks on the same clinical signals, producing analyses from different perspectives, thereby enhancing the model's ability to understand patient states.

In Stage I, we leverage a strong teacher LLM to generate supervision signals under multiple complementary clinical tasks. Each data instance provides the teacher model with the patient state $\mathcal{S}$, relevant clinical guidelines g, and reference physician actions a^*. Task-specific instructions are designed to prompt the teacher to analyze the same patient state from different clinical perspectives.

Specifically, we consider three types of instructions:

- 1) *State Analysis*, which asks the teacher model to synthesize the patient's current hemodynamic status (e.g., blood pressure and heart rate), perfusion indicators (e.g., lactate and urine output), and assess compliance with the clinical definition of septic shock;
- 2) *Patient Dynamics*, which requires the teacher to reason about temporal trends and disease progression, including in-hospital mortality and likely to require vasopressors within the next 24 hours;
- 3) *Decision Making*, which prompts the teacher to conclude with concrete medication levels based on the preceding analysis and the provided clinical guidelines.

For each instruction type q_k, the teacher model produces a task-specific response conditioned on the full clinical context:

$$r_k = \mathrm{TeacherLLM}(\mathcal{S}, g, a^*, q_k), \quad k \in \{\mathrm{analysis}, \mathrm{dynamics}, \mathrm{decision}\}.$$

The final training sample is constructed by concatenating all task responses using a fixed template, forming a unified supervision signal that integrates patient state interpretation, temporal reasoning, and guideline-constrained decision making. These synthesized samples are subsequently used to train the target LLM, enabling it to internalize structured and clinically grounded reasoning patterns demonstrated by the teacher model.

Clinical Guidlines

You act as an AI Clinician. You must strictly follow these rules to analyze the patient state:

1. **Emergency Priority**: Sepsis is a medical emergency. Hypoperfusion requires immediate resuscitation.

2. **Early Resuscitation Rule (0–3h)**: Within the first 3 hours of admission (Hour 0), if there are signs of hypoperfusion, at least **LOW**-level IV fluid is **MANDATORY**.

3. **Vasopressor Threshold**: If MAP $<$ 65 mmHg despite prior fluid resuscitation, Vasopressors **MUST** be initiated or increased.

4. **MAP Target**: For patients with septic shock on vasopressors, the initial target MAP is 65 mmHg rather than higher targets.

5. **Definition of Septic Shock**: A patient is in *Septic Shock* at current time t **ONLY** if **ALL** three conditions are met:
   (a) Vasopressor level > 0 (Currently on vasopressors)
   (b) MAP (meanbp) < 65 mmHg
   (c) Lactate > 2 mmol/L

Figure 4: Clinical guidelines g we used during the data construction process.

Prompt Template for Clinical Decision Generation

Patient Demographics
Age: {{age}} years
Gender: {{gender}}
<*(. . . Other relevant vitals and labs included . . .)*>

Clinical Guidelines for Diagnosis
<Clinical Guidlines>

Doctor's Decision
<Doctor Decision>

Task Instructions
<Task Instructions>

CRITICAL GUARDRAILS
1. **Data Grounding**: Do not hallucinate symptoms not present in the data.
2. **Internal Alignment**: Ensure your reasoning logically leads to the 'Target Action' provided in the reference context (though do not mention you saw the reference).

Please generate the answer directly based on all the provided information, without any additional explanation. Note that you should assume the physician's decision is not given.

Figure 5: Prompt Templates.

F.1.2 *Stage II* TRAINING DATA

Agentic Simulation Reasoning Data For agentic simulation reasoning data, the instruction requires the Teacher LLM to iteratively reason about patient states through interaction with a clinical world model. At each timestep t, the Teacher LLM receives the current patient state $\mathcal{S}_t$, a simulation-oriented instruction q_{sim}, and a predefined simulation plan specifying a small set of candidate treatment actions. The Teacher LLM first generates free-form clinical reasoning and invokes a `simulation` tool to query the world model, which returns predicted future states for the proposed actions:

$$\{\hat{\mathcal{S}}_{t+1}^{(i)}\} = \mathcal{W}(\mathcal{S}_t, \{a_t^{(i)}\})$$

Task Instructions

Instruction of State Analysis
Synthesize the patient's hemodynamic status (BP, HR), perfusion (Lactate, Urine), and check compliance with the 'Septic Shock' definition.

Instruction of Patient Dynamics
Evaluate the trajectory. Is the patient deteriorating? Will they likely need vasopressors or enter shock in the next 24h?

Instruction of Decision Making
Based on the analysis and the Guidelines, conclude with the specific medication levels.
Response Format Requirements
- **Step 1: Chain of Thought**: Write a free-form, detailed clinical note explaining your reasoning. Cite specific data points from the history. (Do not use rigid templates, write as a doctor thinks).
- **Step 2: Final Action**: At the very end, state your decision clearly in the following format:
`ACTION: Vasopressor=[Level], IV Fluid=[Level]`

Figure 6: Task Instructions.

The returned simulated outcomes are then incorporated into subsequent reasoning steps, allowing the Teacher LLM to refine its analysis over multiple simulation rounds before making a final decision. This data teaches the model how to perform lookahead clinical reasoning via explicit Teacher LLM–world model interaction rather than one-shot inference.

Agentic Prescription Data After completing the simulation interactions, the instruction asks the Teacher LLM to commit to a final treatment decision. In this stage, the Teacher LLM receives the patient state $\mathcal{S}_t$, the accumulated simulation feedback from the world model, and a decision-oriented instruction q_{rx}, together with a reference physician action a_t^*. The Teacher LLM is required to generate a clinically grounded rationale and invoke the `prescription` tool to produce a treatment action consistent with the reference:

$$a_t = \text{Teacher LLM}(\mathcal{S}_t, q_{\text{rx}}, \{\hat{\mathcal{S}}_{t+1}^{(i)}\}, a_t^*)$$

This data enables the model to learn how to synthesize world-model-based simulations into a concrete treatment decision aligned with physician practice.

F.1.3 *Stage III* Training Data

In the third stage, we provide the model with the same patient information as in Stage II, construct training prompts in exactly the same format as those used in Stage II, and require the model to complete training through interaction with the world model.

F.2 Training Settings

We employ a three-stage training pipeline consisting of two full-parameter supervised fine-tuning (SFT) stages followed by an agentic reinforcement learning stage with GRPO. All stages are trained on 8 H200 GPUs with mixed-precision bfloat16.

Stage I In the first stage, we perform full-parameter supervised fine-tuning on the backbone model using approximately 100k training instances. The model is trained for 2 epochs with a per-device batch size of 4 and gradient accumulation over 8 steps, resulting in an effective batch size of 256. We use a learning rate of 1×10^{-5} with a warmup ratio of 0.05. The maximum input length is set to 81,920 tokens to support long-context clinical reasoning.

Stage II The second stage continues full-parameter supervised fine-tuning with a smaller, higher-quality dataset of approximately 1k instances, focusing on refining decision-making patterns. The

training setup remains largely consistent with Stage I, except that the per-device training batch size is reduced to 1 to accommodate more complex samples. The model is trained for 2 epochs with the same learning rate (1×10^{-5}), global batch size (64), warmup ratio (0.05), DeepSpeed ZeRO Stage 2, and maximum sequence length (81,920 tokens).

Stage III In the final stage, we apply Group Relative Policy Optimization (GRPO) Shao et al. (2024) using approximately 3k training trajectories. The agent interacts with a clinical world model environment and is trained on 8 GPUs. We use a learning rate of 1×10^{-6} with a warmup ratio of 0.05. The training batch size is 16, with a PPO mini-batch size of 8 and a micro-batch size of 1 per GPU. The maximum prompt and response lengths are set to 8,192 and 81,920 tokens, respectively, with rollout model length up to 102,400 tokens to support long-horizon multi-turn interactions. KL regularization is applied with a coefficient of 0.001. We terminate the GRPO training after 300 steps.

F.3 Reward Function Details

We define the rollout reward to encourage long-term survival, short-term physiological stabilization, and guideline-consistent treatment. As described in Section 4.4, the total reward is

$$R(\tau) = R_{\text{out}}(s_T) + \sum_{t=0}^{T-1} r(s_t, s_{t+1}) - \lambda_g P_g(\tau).$$

Following the intermediate reward design used in DDQN for sepsis treatment Raghu et al. (2017), the step-wise physiological reward is defined as:

$$\begin{aligned} r(s_t, s_{t+1}) = C_0 \mathbb{I}\left(s_{t+1}^{\text{SOFA}} = s_t^{\text{SOFA}} \wedge s_{t+1}^{\text{SOFA}} > 0\right) \\ + C_1 \left(s_{t+1}^{\text{SOFA}} - s_t^{\text{SOFA}}\right) \\ + C_2 \tanh\left(s_{t+1}^{\text{Lactate}} - s_t^{\text{Lactate}}\right), \end{aligned} \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function. We set $C_0 = -0.025$, $C_1 = -0.125$, and $C_2 = -2$. The first term penalizes stagnation in nonzero SOFA states, the second term penalizes SOFA deterioration, and the third term penalizes lactate increase. The hyperbolic tangent caps the magnitude of the lactate term, preventing intermediate rewards from dominating terminal outcomes.

For terminal outcomes, we assign $R_{\text{out}} = +15$ for survival and $R_{\text{out}} = -15$ for mortality. To discourage clinically implausible actions, we apply a fixed penalty of -10 for sepsis guideline violations. If the agent fails to reach a terminal decision within the maximum allowed interaction steps and repeatedly queries the world model, we apply an additional penalty of -5. For training stability, the total reward is scaled by 0.1 and clipped to $[-2, 2]$.

F.4 Reproducibility and Evaluation Protocol

Table 11 summarizes the data allocation and shared comparison settings for the main sepsis experiments. The ICU-stay counts describe source cohorts, whereas the training-instance and trajectory counts in Appendix F.2 describe examples constructed from those stays.

Table 11: Data allocation and comparison protocol for the main sepsis experiments.

Component	Protocol
Dataset split	The 20,092 MIMIC-IV ICU stays are split into training, validation, and test sets at a 7:2:1 ratio. Clinical observations are aggregated into 4-hour steps.
World model	The Clinical World Model is trained on the full 70% training split and then frozen before agent training.
Agent stages	The training split is partitioned at the ICU-stay level into disjoint subsets of approximately 11,000, 1,000, and 2,000 stays for Stages I, II, and III, respectively.
RL baselines	DDQN, AI Clinician, and WD3QNE follow their released implementations and use the same 42-variable state space, 5×5 fluid–vasopressor action space, and 4-hour decision interval as SepsisAgent.
LLM comparison	World-model-augmented LLMs and SepsisAgent use the same Clinical World Model, prompt, and propose–simulate–refine workflow.
Evaluation cohort	The main treatment-policy comparisons use the same 725 held-out MIMIC-IV test episodes for RL and LLM methods.

The resulting Stage I and II training sets contain approximately 100k and 1k constructed instances, respectively; Stage III uses approximately 3k training trajectories. Detailed training hyperparameters are given in Appendix F.2, and the OPE estimators are defined in Appendix G.

G Off-Policy Evaluation Details

We evaluate treatment policies using off-policy evaluation (OPE), since direct online interaction with ICU patients is infeasible. Let $\mathcal{D} = \{\tau_i\}_{i=1}^{N}$ denote a set of retrospective trajectories collected under the behavior policy π_b, where each trajectory is

$$\tau_i = \{(s_{i,0}, a_{i,0}, r_{i,0}), \ldots, (s_{i,T_i}, a_{i,T_i}, r_{i,T_i})\}.$$

The target policy to be evaluated is denoted by π_e. We define the cumulative return of trajectory i as

$$G_i = \sum_{t=0}^{T_i} \gamma^t r_{i,t},$$

where γ is the discount factor.

Importance ratios. For each trajectory, the trajectory-level and per-decision importance ratios are defined as

$$\rho_i = \prod_{t=0}^{T_i} \frac{\pi_e(a_{i,t} \mid s_{i,t})}{\pi_b(a_{i,t} \mid s_{i,t})}, \qquad \rho_{i,t} = \prod_{k=0}^{t} \frac{\pi_e(a_{i,k} \mid s_{i,k})}{\pi_b(a_{i,k} \mid s_{i,k})}.$$

These ratios correct the distribution mismatch between the clinician behavior policy and the evaluated policy.

Weighted Importance Sampling. Weighted Importance Sampling (WIS) normalizes trajectory-level importance weights to reduce variance Precup et al. (2000):

$$\hat{V}_{\mathrm{WIS}} = \frac{\sum_{i=1}^{N} \rho_i G_i}{\sum_{i=1}^{N} \rho_i + \epsilon},$$

where ϵ is a small constant for numerical stability.

Weighted Per-Decision Importance Sampling. Weighted Per-Decision Importance Sampling (WPDIS) applies normalization at each decision step rather than at the full-trajectory level Precup et al. (2000):

$$\hat{V}_{\mathrm{WPDIS}} = \sum_{t=0}^{T} \gamma^t \frac{\sum_{i=1}^{N} \mathbb{I}(t \le T_i) \rho_{i,t} r_{i,t}}{\sum_{i=1}^{N} \mathbb{I}(t \le T_i) \rho_{i,t} + \epsilon}.$$

Compared with WIS, WPDIS better accounts for step-wise treatment effects and can reduce variance in long clinical trajectories.

Doubly Robust estimation. We also report the Doubly Robust (DR) estimator Jiang & Li (2016), which combines importance weighting with an approximate value model. Let $\hat{Q}(s, a)$ be the estimated action-value function and

$$\hat{V}(s) = \sum_{a \in \mathcal{A}} \pi_e(a \mid s) \hat{Q}(s, a).$$

The per-trajectory DR estimate is computed recursively as

$$\hat{V}_{i,t}^{\mathrm{DR}} = \hat{V}(s_{i,t}) + \rho_{i,t} \left(r_{i,t} + \gamma \hat{V}_{i,t+1}^{\mathrm{DR}} - \hat{Q}(s_{i,t}, a_{i,t}) \right),$$

with terminal condition $\hat{V}_{i,T_i+1}^{\mathrm{DR}} = 0$. The final DR estimate is

$$\hat{V}_{\mathrm{DR}} = \frac{1}{N} \sum_{i=1}^{N} \hat{V}_{i,0}^{\mathrm{DR}}.$$

The accuracy of DR depends on the quality of the estimated value function. Prior work on WD3QNE notes that inaccurate target Q-value estimation can introduce overestimation or underestimation bias, and proposes an adaptive weighted target Q-value function to balance the overestimation tendency of Dueling DQN and the underestimation tendency of D3QN Wu et al. (2023). Therefore, we follow the WD3QNE implementation to estimate $\hat{Q}(s, a)$ for DR evaluation.

H Cross-Dataset Evaluation

We assess generalization beyond the MIMIC-IV test set in two complementary settings. MIMIC-III provides a cross-period evaluation from the same hospital, whereas the multicenter eICU dataset probes transfer across institutions. We report retrospective off-policy evaluation (OPE) in both settings.

Cross-period evaluation on MIMIC-III. We evaluate SepsisAgent on 2,862 MIMIC-III episodes. As shown in Table 12, SepsisAgent achieves the highest DR and WIS among the compared methods and remains competitive on WPDIS.

Table 12: Cross-period OPE on 2,862 MIMIC-III episodes.

Method	**DR** (↑)	**WIS** (↑)	**WPDIS** (↑)
Clinician	5.90	5.77	14.92
AI Clinician	<u>9.91</u>	6.39	14.25
DDQN	7.03	6.03	14.56
WD3QNE	6.40	<u>9.16</u>	**18.67**
SepsisAgent	**10.08**	**11.50**	<u>16.22</u>

Cross-institution evaluation on eICU. We further evaluate the policies on the multicenter eICU dataset. As shown in Table 13, SepsisAgent achieves strong performance across all three OPE estimators, obtaining the highest DR and WPDIS among the compared methods and exceeding the clinician reference on DR, WIS, and WPDIS.

Table 13: Cross-institution OPE on the multicenter eICU dataset.

Method	**DR** (↑)	**WIS** (↑)	**WPDIS** (↑)
Clinician	5.24	5.74	12.95
WD3QNE	7.57	**8.79**	13.77
o3	5.79	5.21	13.14
o3 + WM	<u>8.47</u>	6.08	<u>18.91</u>
Qwen3-4B	5.47	5.35	10.79
SepsisAgent	**8.99**	<u>7.65</u>	**20.32**

I FAILURE MODE ANALYSIS

This section complements Section 3.3 by analyzing why world-model access alone may fail. We find that the main failure mode of generic LLMs with world-model feedback is a greedy and aggressive treatment strategy: the model over-follows short-horizon simulated improvements, such as immediate MAP recovery, while overlooking longer-term risks such as excessive vasopressor or fluid exposure.

Patient Status (stay_id=33066642, Current t=6):
- **MAP Trajectory (last 24h):** 54 → 65 → 60 → 65 → 67 → 81 → 60 mmHg (¡ 65, Hypotensive).
- **Lactate Trajectory:** 3.9 → 3.3 → 3.1 → 3.1 → 3.1 → 1.5 → 1.5 mmol/L.
- **SOFA Score Trajectory:** 10 → 9 → 8 → 6 → 6 → 8 → 8 (Hidden to LLM, requires inference).
- **Medication History [Vaso, Fluid]:** [None, V.High] → [Med, V.High] → [High, V.High] → [High, V.High] → [High, V.High] → [High, V.High].

World Model Simulations:
- *Action A [High, V.High]:* Predicted MAP → **65** mmHg.
- *Action B [Low, Med]:* Predicted MAP → **63** mmHg.

Decision Comparison:
- **o3 + WM:** Selects *Action A*.
  Reasoning: "Simulation confirms Action A hits the MAP target (65), whereas Action B fails (63)." → **Outcome: Shock in 24h → Died.**
- **SepsisAgent:** Selects *Action B* (Matches Clinician).
  Reasoning: "Lactate is stable (1.5). Patient has massive fluid load. Continuing V.High fluids risks overload. I will taper to [Low, Med] despite MAP 63." → **Outcome: Tapered → Survived.**

Figure 7: SepsisAgent prioritizes long-term survival (avoiding overload) over short-term metrics.(Med = Medium, V. = Very)

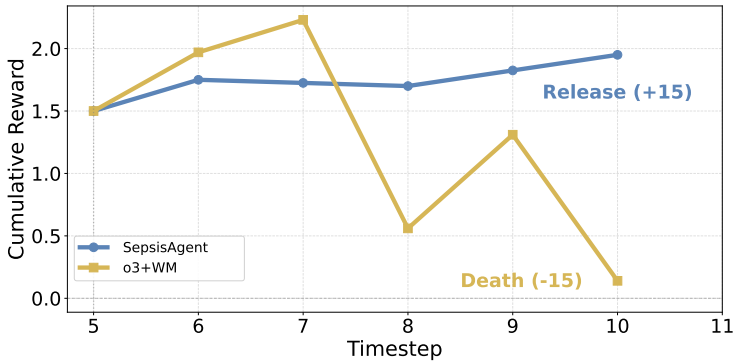


Figure 8: Cumulative reward comparison between SepsisAgent and o3 + WM in the representative failure case.

We further quantify this pattern using the expert-defined unsafe-action metrics. We sample 241 episodes and ask four human experts to manually validate the identified failures, including two medical doctors and two medical master's degree holders, all from emergency medicine. Among world-model-augmented LLMs, **o3 + WM** exhibits overdosing failures in 7.1% of episodes, **Gemini-3-Flash + WM** in 7.8%, and **GPT-4.1-mini + WM** in 10.3%. In contrast, **SepsisAgent** shows only 0.8% overdosing failures, corresponding to 2 cases among 241 episodes. These results support the observation that world-model access alone is insufficient; the agent must learn to interpret simulated patient responses in the context of long-term clinical risk.

J UNCERTAINTY AND STATISTICAL TESTING FOR OFF-POLICY EVALUATION

The OPE point estimates in the main results and Appendix H do not describe their sampling variability. We therefore use 2,000 patient-level paired bootstrap resamples to obtain 95% confidence intervals for DR, WIS, and WPDIS on the MIMIC-IV test set and the cross-institution eICU dataset. Each paired comparison evaluates SepsisAgent and the clinician reference on the same resampled

patients. Table 14 reports the resulting intervals and the corresponding p-values for the SepsisAgent–clinician comparison.

Table 14: OPE estimates with 95% bootstrap confidence intervals. p-values compare SepsisAgent with the clinician reference using 2,000 patient-level paired bootstrap resamples.

Metric	Clinician (95% CI)	SepsisAgent (95% CI)	p-value
	MIMIC-IV		
DR	5.06 [4.81, 5.25]	10.01 [8.31, 11.56]	< 0.001
WIS	5.27 [5.00, 5.77]	11.14 [6.26, 14.78]	< 0.001
WPDIS	10.82 [8.60, 12.50]	23.40 [12.70, 29.93]	< 0.01
	eICU		
DR	5.24 [4.95, 6.86]	8.99 [7.37, 11.01]	< 0.001
WIS	5.74 [5.33, 7.26]	7.65 [7.26, 11.25]	$= 0.05$
WPDIS	12.95 [11.39, 14.25]	20.32 [14.44, 25.96]	< 0.001

The paired comparisons favor SepsisAgent under all three estimators on MIMIC-IV. On eICU, the DR and WPDIS comparisons also favor SepsisAgent, whereas the WIS comparison is borderline ($p = 0.05$). These tests quantify sampling uncertainty for the clinician comparison.

K Example Reasoning Trace

This section presents a real reasoning process of SepsisAgent (shown in Figure K). As illustrated, SepsisAgent invokes tools during inference, repeatedly performing simulations to validate its hypotheses and analyzing the resulting outcomes to arrive at the final conclusion.

A reasoning process of SepsisAgent (Some repetitive reasoning steps in the middle have been omitted due to the extensive length.)

```
system
You are a helpful assistant.

# Tools

You may call one or more functions to assist with the user query.

You are provided with function signatures within <tools></tools>
    XML tags:
<tools>
{"type": "function", "function": {"name": "simulation",
    "description": "Simulate patient outcomes for different
    treatment actions before making a final decision. Use this when
    you want to compare multiple treatment options.", "parameters":
    {"type": "object", "properties": {"actions": {"type": "array",
    "description": "List of treatment actions to simulate. Each
    action is '[vasopressor_level, iv_fluid_level]' where levels
    are 0-4. Maximum 3 actions per call.", "items": {"type":
    "string"}}}, "required": ["actions"]}}}
{"type": "function", "function": {"name": "prescription",
    "description": "Execute the final treatment decision. Use this
    when you are confident about the best treatment after analysis
    or simulation.", "parameters": {"type": "object", "properties":
    {"vasopressor": {"type": "integer", "description": "Vasopressor
    level (0-4): None(0), Low(1), Medium(2), High(3), Very
    High(4)"}, "iv_fluid": {"type": "integer", "description": "IV
    Fluid level (0-4): None(0), Low(1), Medium(2), High(3), Very
    High(4)"}}, "required": ["vasopressor", "iv_fluid"]}}}
```

```
</tools>

For each function call, return a json object with function name
    and arguments within <tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>
user
You are an expert ICU physician AI assistant specializing in
    sepsis treatment decisions.

## Patient Information
- Age: 76 years
- Gender: Male
- Charlson Comorbidity Index: 3.0

## Important Notes
- Patient vital signs and lab values are monitored every 4 hours
- You will receive updates on patient status and make treatment
    decisions accordingly
- Treatment actions take effect over the next 4-hour period
- You are starting from ICU admission time (t=0)

## Treatment Levels
- IV Fluid: None (0), Low (1), Medium (2), High (3), Very High (4)
- Vasopressor: None (0), Low (1), Medium (2), High (3), Very High
    (4)

## Available Tools
You have access to two tools:

1. **simulation**: Simulate patient outcomes for different
    treatment actions before making a final decision.
- Parameter: actions (list of "[vasopressor_level,
    iv_fluid_level]" strings, max 3 actions)

2. **prescription**: Execute the final treatment decision.
- Parameters: vasopressor (int 0-4), iv_fluid (int 0-4)

## Clinical Protocols (Strict Adherence Required)
1. **Emergency Priority**: Sepsis is a medical emergency.
    Hypoperfusion requires immediate resuscitation.
2. **Early Resuscitation Rule (0-3h)**: Within the first 3 hours
    of admission (Hour 0), if there are signs of hypoperfusion, at
    least LOW-level IV fluid is MANDATORY.
3. **Vasopressor Threshold**: If MAP remains < 65 mmHg after
    adequate fluid resuscitation, vasopressor support should be
    considered.
4. **MAP Target**: For patients with septic shock on vasopressors,
    the initial target MAP is 65 mmHg rather than higher targets.
5. **Definition of Septic Shock**: A patient is in 'Septic Shock'
    ONLY if ALL three conditions are met:
(a) Vasopressor level > 0 (Currently on vasopressors)
(b) MAP (meanbp) < 65 mmHg
(c) Lactate > 2 mmol/L
# Hour 0 Since ICU Admission (timestep t=0)

## Vital Signs History
- heart_rate(bpm): [63.9]
- sysbp(mmHg): [107.8]
- diabp(mmHg): [58.5]
- meanbp(mmHg): [69.8]
- resp_rate(breaths/min): [21.4]
```

```
- spo2(%): [91.2]
- temp_c(C): [37.0]
- fio2(%): [100.0]

## Laboratory Values History
- albumin(g/dL): [3.1]
- alt(U/L): [19.0]
- ast(U/L): [30.0]
- base_excess(mEq/L): [4.0]
- bicarbonate(mEq/L): [21.0]
- bilirubin_total(mg/dL): [0.6]
- bun(mg/dL): [21.0]
- calcium_free(mmol/L): [1.1]
- calcium_total(mg/dL): [8.0]
- chloride(mEq/L): [102.0]
- creatinine(mg/dL): [1.1]
- glucose(mg/dL): [217.0]
- hemoglobin(g/dL): [12.1]
- hematocrit(%): [36.2]
- lactate(mmol/L): [2.6]
- magnesium(mEq/L): [2.2]
- paco2(mmHg): [34.0]
- pao2(mmHg): [107.0]
- ph(): [7.4]
- platelet(K/uL): [151.0]
- potassium(mEq/L): [4.1]
- pt(sec): [12.9]
- ptt(sec): [22.1]
- sodium(mEq/L): [134.0]
- total_co2(mEq/L): [23.0]
- wbc(K/uL): [9.1]

## Urine Output History
- output_4hourly(mL/4h): [600.0]

---
You may call `simulation` to predict outcomes, or call
    `prescription` to make your final decision.
assistant
This is a 76-year-old male with a Charlson Comorbidity Index of
    3.0, admitted to the ICU for sepsis management. On initial
    assessment, his mean arterial pressure (MAP) is 69.8 mmHg,
    which is above the critical threshold of 65 mmHg, and his
    lactate is mildly elevated at 2.6 mmol/L, suggesting some
    degree of tissue hypoperfusion. His urine output over the last
    4 hours is robust at 600 mL, indicating adequate renal
    perfusion at this time. He is not currently on vasopressor
    support, and his MAP is above 65 mmHg, so he does not meet
    criteria for septic shock at this time.

Given the mild elevation in lactate and the absence of shock,
    early fluid resuscitation is warranted, but vasopressor support
    is not indicated at this moment. The patient's overall
    prognosis appears favorable, with no evidence of multi-organ
    failure or refractory shock. Based on the current trajectory,
    it is unlikely that he will require vasopressors or develop
    septic shock within the next 24 hours, and his 90-day mortality
    risk is low.

To guide initial management, I will simulate three reasonable
    fluid strategies: low, medium, and very high IV fluid
    administration, all without vasopressor support, to assess
    their impact on MAP, lactate, and overall stability.
```

<tool_call>
{"name": "simulation", "arguments": {"actions": ["[0,1]", "[0,2]", "[0,4]"]}}
</tool_call>
user
<tool_response>
{"result": "## Simulation Results\n\n### Option 1: IV Fluid=Low, Vasopressor=None\nPredicted patient state after 4 hours:\n\n**Vital Signs:**\n- Heart Rate: 50.9 bpm\n- Sysbp: 113.4 mmHg\n- Diabp: 61.2 mmHg\n- Meanbp: 70.3 mmHg\n- Resp Rate: 20.1 breaths/min\n- Spo2: 99.5 %\n- Temp C: 36.8 \u00b0C\n\n**Key Lab Values:**\n- Lactate: 2.3 mmol/L\n- Creatinine: 1.5 mg/dL\n- Bilirubin Total: 0.7 mg/dL\n- Platelet: 58.8 K/\u03bcL\n- Ph: 7.3 \n- Pao2: 103.3 mmHg\n\n**Urine Output:** 55.6 mL/4h\n\n### Option 2: IV Fluid=Medium, Vasopressor=None\nPredicted patient state after 4 hours:\n\n**Vital Signs:**\n- Heart Rate: 60.6 bpm\n- Sysbp: 104.8 mmHg\n- Diabp: 38.8 mmHg\n- Meanbp: 80.7 mmHg\n- Resp Rate: 20.8 breaths/min\n- Spo2: 91.3 %\n- Temp C: 36.8 \u00b0C\n\n**Key Lab Values:**\n- Lactate: 2.8 mmol/L\n- Creatinine: 1.6 mg/dL\n- Bilirubin Total: 1.6 mg/dL\n- Platelet: 220.3 K/\u03bcL\n- Ph: 7.4 \n- Pao2: 127.8 mmHg\n\n**Urine Output:** 1719.3 mL/4h\n\n### Option 3: IV Fluid=Very High, Vasopressor=None\nPredicted patient state after 4 hours:\n\n**Vital Signs:**\n- Heart Rate: 81.6 bpm\n- Sysbp: 118.0 mmHg\n- Diabp: 63.0 mmHg\n- Meanbp: 82.6 mmHg\n- Resp Rate: 18.3 breaths/min\n- Spo2: 90.2 %\n- Temp C: 36.9 \u00b0C\n\n**Key Lab Values:**\n- Lactate: 1.3 mmol/L\n- Creatinine: 0.7 mg/dL\n- Bilirubin Total: 1.1 mg/dL\n- Platelet: 180.5 K/\u03bcL\n- Ph: 7.5 \n- Pao2: 93.7 mmHg\n\n**Urine Output:** 679.3 mL/4h\n\n---\nYou may call `simulation` again, or call `prescription` to make your final decision."}
</tool_response>
assistant
The patient's initial presentation is notable for a MAP of 69.8 mmHg, which is above the critical threshold for shock, and a lactate of 2.6 mmol/L, indicating mild tissue hypoperfusion. He is not on vasopressors, and his urine output is robust at 600 mL over the last 4 hours, suggesting adequate renal perfusion. There is no evidence of multi-organ failure or refractory shock at this time.

Simulation of low, medium, and very high IV fluid strategies shows that all options maintain MAP above 65 mmHg, but the very high fluid strategy results in the most favorable improvement in lactate (dropping to 1.3 mmol/L) and MAP (rising to 82.6 mmHg), with good urine output (679.3 mL/4h). Importantly, there is no indication for vasopressor support at this time, and the patient is not in septic shock.

Given the mild lactate elevation and the absence of shock, aggressive fluid resuscitation is appropriate to optimize perfusion and clear lactate, especially in the early phase of sepsis management. The patient's overall prognosis remains favorable, and there is no evidence to suggest he will require vasopressors or develop shock in the next 24 hours.

I will proceed with very high IV fluid administration and no vasopressor support.
<tool_call>
{"name": "prescription", "arguments": {"vasopressor": 0, "iv_fluid": 4}}

```
</tool_call>
user
<tool_response>
{"result": "Based on your decision, the patient received Very High
    IV fluid over the past 4 hours.\n\n# Hour 4 Since ICU
    Admission\n\n## Vital Signs History\n- heart_rate(bpm): [63.9,
    53.7]\n- sysbp(mmHg): [107.8, 103.2]\n- diabp(mmHg): [58.5,
    58.3]\n- meanbp(mmHg): [69.7, 77.1]\n- resp_rate(breaths/min):
    [21.4, 24.2]\n- spo2(%): [91.2, 91.2]\n- temp_c(\u00b0C):
    [37.0, 37.1]\n- fio2(%): [100.0, 43.4]\n\n## Laboratory Values
    History\n- albumin(g/dL): [3.1, 3.3]\n- alt(U/L): [19.0,
    66.9]\n- ast(U/L): [30.0, -10.7]\n- base_excess(mEq/L): [4.0,
    4.6]\n- bicarbonate(mEq/L): [21.0, 22.1]\n-
    bilirubin_total(mg/dL): [0.6, 0.8]\n- bun(mg/dL): [21.0,
    15.2]\n- calcium_free(mmol/L): [1.1, 1.1]\n-
    calcium_total(mg/dL): [8.0, 8.4]\n- chloride(mEq/L): [102.0,
    102.8]\n- creatinine(mg/dL): [1.1, 0.9]\n- glucose(mg/dL):
    [217.0, 190.5]\n- hemoglobin(g/dL): [12.1, 12.0]\n-
    hematocrit(%): [36.2, 36.9]\n- lactate(mmol/L): [2.6, 1.8]\n-
    magnesium(mEq/L): [2.2, 2.5]\n- paco2(mmHg): [34.0, 26.4]\n-
    pao2(mmHg): [107.0, 132.1]\n- ph(): [7.4, 7.4]\n-
    platelet(K/\u03bcL): [151.0, 130.8]\n- potassium(mEq/L): [4.1,
    3.9]\n- pt(sec): [12.9, 12.4]\n- ptt(sec): [22.1, 37.4]\n-
    sodium(mEq/L): [134.0, 135.9]\n- total_co2(mEq/L): [23.0,
    21.5]\n- wbc(K/\u03bcL): [9.1, 11.4]\n\n## Urine Output\n-
    output_4hourly(mL/4h): [600.0, 701.4]\n\n## Treatment
    History\n- Hour 4: Vasopressor=None, IV Fluid=Very
    High\n\n---\nYou may call `simulation` to predict outcomes (up
    to 3 more round(s)), or call `prescription` to make your final
    decision."}
</tool_response>
assistant
At hour 4, the patient has received very high IV fluid
    resuscitation. His MAP has improved from 69.7 to 77.1 mmHg, and
    lactate has decreased from 2.6 to 1.8 mmol/L, indicating
    effective resuscitation and improved tissue perfusion. Urine
    output remains strong at 701.4 mL over the last 4 hours, and
    there is no evidence of acute kidney injury (creatinine has
    decreased to 0.9 mg/dL). Other labs show mild increases in
    liver enzymes, but no significant organ dysfunction. The
    patient remains afebrile, and oxygenation is stable.

He is not on vasopressor support, and his MAP is well above 65
    mmHg, so he does not meet criteria for septic shock. Given the
    improvement in lactate and MAP, there is no indication for
    vasopressor initiation at this time. The patient's overall
    trajectory is favorable, and there is no evidence to suggest he
    will require vasopressors or develop shock in the next 24 hours.

To guide ongoing management, I will simulate three options:
    continuing very high IV fluid, stepping down to medium IV
    fluid, or adding low-dose vasopressor with medium IV fluid, to
    assess the best approach for maintaining stability and avoiding
    fluid overload.
<tool_call>
{"name": "simulation", "arguments": {"actions": ["[0,4]", "[0,2]",
    "[1,2]"]}}
</tool_call>
user

[[... (Omit some steps here)]]

assistant
```

```
At hour 48, the patient remains stable with a MAP of 86.5 mmHg,
    lactate of 1.2 mmol/L, and urine output of 184.7 mL/4h.
    Creatinine is 1.6 mg/dL, and there is no evidence of acute
    kidney injury or multi-organ dysfunction. The patient is not on
    vasopressor support, and his MAP and lactate are well
    controlled, so he does not meet criteria for septic shock.

Simulation of continued very high IV fluid, medium IV fluid, or
    low-dose vasopressor with medium IV fluid all predict stable or
    improved hemodynamics, but there is no indication for
    vasopressor initiation at this time. The patient's MAP is well
    above 65 mmHg, and lactate is not persistently elevated, so
    shock is not imminent.

Given the patient's current stability, ongoing aggressive fluid
    resuscitation is appropriate to maintain perfusion and lactate
    clearance, especially in the early phase of sepsis management.
    There is no indication for vasopressor support, and the risk of
    shock or vasopressor requirement in the next 24 hours remains
    low.

I will continue very high IV fluid administration and no
    vasopressor support.
<tool_call>
{"name": "prescription", "arguments": {"vasopressor": 0,
    "iv_fluid": 4}}
</tool_call>
user
<tool_response>
{"result": "Based on your decision, the patient received Very High
    IV fluid over the past 4 hours.\n\n## Patient Status Update
    (Hour 52)\nGood news! The patient's condition has stabilized
    sufficiently for ICU discharge."}
</tool_response>
```