

MedACE: A Time-Aware Asynchronous Clinical Environment for Long-Horizon Medical Treatment Decision-Making

Zhiting Fan¹, Ruizhe Chen¹, Hualiang Wang², Ru Peng¹,
Zijie Meng¹, Wen MA¹, Jian Wu¹, Zuozhu Liu¹
¹Zhejiang University ²ByteDance

Abstract

Recent advances in Medical Large Language Models have shown strong performance in medical question answering and clinical knowledge understanding, yet real-world clinical practice requires long-horizon, dynamic treatment decision-making rather than static prediction. Existing medical agent environments are mostly turn-based, assuming patient states change after actions while overlooking natural disease progression and delayed treatment effects. We propose **MedACE**, a time-aware asynchronous clinical world model, where patient evolution is formulated as an elapsed-time-conditioned process driven by historical observations, treatment actions, clinical context, and disease-level knowledge. Built from time-stamped MIMIC-IV trajectories, our JEPA-style simulator, **MedWorld-JEPA**, learns future patient state representations in latent space rather than directly generating noisy clinical observations. Based on this simulator, we construct an asynchronous treatment environment with global time advancement, closed-loop patient simulation, multi-step physician-agent interaction, and outcome-grounded rewards. Experiments on Hospital and ICU scenarios demonstrate competitive physiological forecasting, strong clinical consistency, and the importance of jointly modeling time, disease progression, and treatment effects. Agent evaluations further show that current LLM-based physicians struggle with long-horizon treatment decisions, highlighting our environment as a challenging foundation for future medical treatment agents.

1 Introduction

Recent years have witnessed rapid progress in Medical Large Language Models (Med-LLMs), which have achieved strong performance in medical question answering, clinical knowledge understanding, and static case analysis (Lee et al., 2020; Li et al., 2023; Zhang et al., 2023; Sellergren et al., 2025;

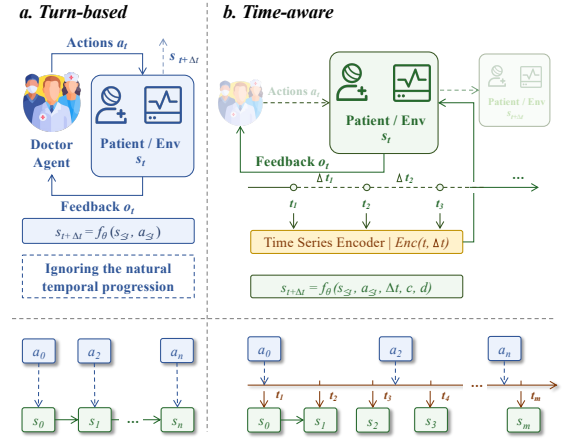


Figure 1: Differences between action-driven turn-based and time-aware asynchronous clinical environments.

Jiang et al., 2025a). However, real-world clinical practice is not a static benchmark, but a long-horizon interactive process requiring continuous observation, dynamic adjustment, and sequential decision-making (Li et al., 2024; Zhu et al., 2026). This calls for dynamic clinical environments that can support long-term interaction, feedback, and evaluation of medical agents, similar to the interactive environments that have enabled recent progress in general-purpose LLM agents (Liu et al., 2025).

Building such environments in medicine is challenging due to costly clinical interaction, long treatment cycles, patient safety concerns, and the difficulty of large-scale repeated experiments. Existing studies have explored LLM-simulated patients for doctor-patient conversations (Qiu et al., 2025; Johri et al., 2025; Zhao et al., 2026; Schmidgall et al., 2024) and virtual EHR platforms for clinical data processing (Shi et al., 2024; Jiang et al., 2025b; Xu et al., 2025). However, treatment decision-making is more complex than medical questioning or information extraction, as it requires long-horizon reasoning over time-sensitive, heterogeneous patient trajectories and complex state transi-

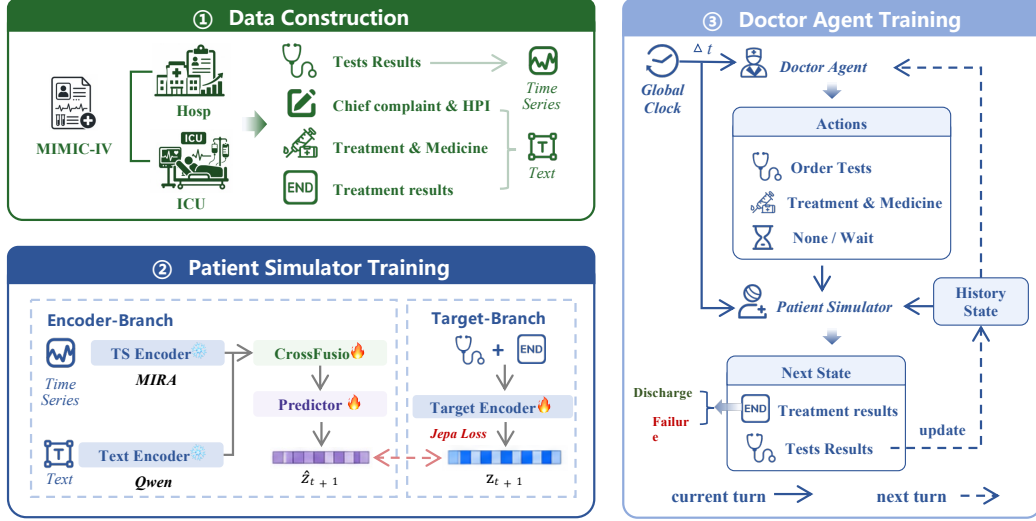


Figure 2: (1) **Data Construction:** Timestamped MIMIC-IV trajectories are organized into structured physiological time series and textual clinical context under Hospital and ICU scenarios. (2) **Patient Simulator Training:** A dual-branch JEPA framework jointly models clinical time series and textual semantics to predict future latent patient states. (3) **Doctor Agent Training:** Based on the patient simulator, we construct an interactive treatment environment where doctor agents interact with asynchronously evolving patient states through long horizon decisions.

tions. Existing treatment-oriented simulators still largely follow a turn-based paradigm, where patient states change mainly after physician actions (Wu et al., 2026; Choudhary et al., 2024). This passive-response assumption overlooks continuous disease progression and delayed treatment effects, making it difficult for agents to learn intervention timing, delayed-action costs, and long-term treatment consequences.

To address these limitations, we model clinical environments as time-aware dynamical systems jointly driven by natural disease progression and clinical interventions. Instead of assuming that state transitions are triggered only by discrete physician actions, we formulate future patient states as:

$$s_{t+\Delta t} = f_{\theta}(s_{\leq t}, a_{\leq t}, \Delta t, c, d), \quad (1)$$

where $s_{\leq t}$ denotes the historical patient trajectory, $a_{\leq t}$ denotes historical treatment actions, Δt represents elapsed time, c denotes textual clinical context, and d denotes disease-level knowledge. Under this formulation, future patient states are determined by disease progression, treatment history, elapsed time, and clinical semantic context, rather than by a single-step action alone.

Motivated by this formulation, we propose **MedWorld-JEPA**, a time-aware clinical world model for treatment-oriented medical agents. The idea is to model patient evolution as an elapsed-time-conditioned latent dynamics process, rather

than a turn-based response triggered only by physician actions. We introduce a JEPA-style (Chen et al., 2025) latent patient simulator that learns clinical state representations from timestamped patient trajectories, providing an interface for modeling sparse, noisy, and heterogeneous clinical observations in latent space.

Built upon this world model, we construct **MedACE**, an asynchronous treatment environment for long-horizon physician-agent interaction. **MedACE** contains Hosp and ICU scenarios, where Hosp emphasizes longer-term inpatient progression and complex examinations, while ICU focuses on dense physiological time series and urgent clinical conditions. Following the JEPA formulation, we denote it as a clinical world model because it predicts future latent patient states conditioned on historical observations and actions. The environment maintains latent patient states, advances a clinical clock, and decodes predicted latent states into observable clinical feedback when needed. Physician agents interact with the evolving patient trajectory through a closed loop of *observe* $\rightarrow$ *act* $\rightarrow$ *simulate* $\rightarrow$ *observe*.

Experiments on Hospital and ICU scenarios demonstrate the effectiveness of the **MedACE** framework. The patient simulator achieves competitive physiological forecasting and strong clinical consistency compared with time-series and

Table 1: Sequence distribution statistics on the MIMIC-IV dataset.

Setting	Patients	Stays	LOS/stay (h)	Obs./stay	Actions/stay	Obs./h	Actions/h
Hosp	223,452	546,028	125.15	180.95	75.25	1.55	0.50
ICU	65,366	94,458	1.44	3720.49	1148.12	87.00	39.98

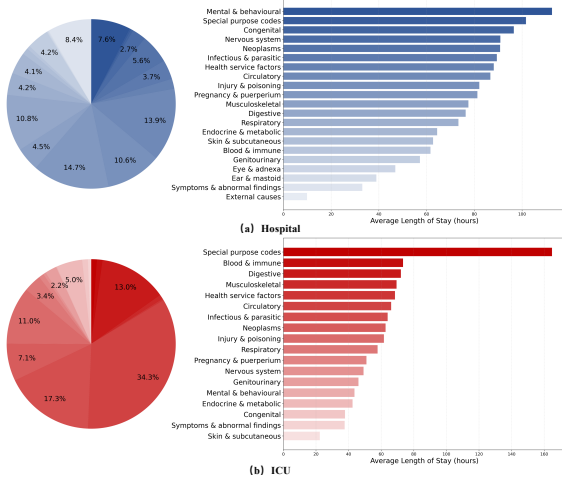


Figure 3: Disease distribution and average clinical stay duration under Hospital and ICU scenarios.

LLM-based baselines. Ablation studies validate the necessity of jointly modeling elapsed time, natural disease progression, and action-conditioned treatment effects. Moreover, evaluations of LLM-based physician agents reveal that current general-purpose LLMs still struggle with long-horizon treatment decision-making under high-frequency physiological changes and delayed feedback. Our main contributions are summarized as follows:

- We formulate clinical state transition as a time-aware asynchronous process, modeling patient evolution as the joint effect of disease progression, treatment actions, elapsed time, clinical context, and medical knowledge.
- We construct a latent clinical world model from timestamped trajectories, where a JEPA-style simulator learns future patient state representations instead of directly generating noisy and irregular clinical observations.
- We build an asynchronous treatment environment with time progression, closed-loop patient simulation, and outcome-grounded rewards, enabling evaluation of agents on both action selection and intervention timing.

2 Method

2.1 Time-Aware Trajectory Construction

We construct timestamped clinical trajectories from MIMIC-IV to support long-horizon and asynchronously evolving patient state modeling. Each trajectory contains structured physiological observations, treatment actions, and textual clinical context, enabling future patient evolution to be modeled under both disease progression and clinical intervention history. To capture different temporal dynamics, we build two scenarios: a Hospital scenario for general inpatients and an ICU scenario for critically ill patients with higher-frequency physiological changes. We split the dataset at the patient level to avoid trajectory leakage. Table 1 summarizes the constructed datasets.

We generate supervised samples using a sliding temporal window strategy. Given a reference time t , the model observes historical information within $[t - H, t]$ and predicts future patient states in $(t, t + \Delta]$. The history length H and prediction horizon Δ are set differently for Hospital and ICU scenarios to reflect their distinct clinical temporal resolutions. Structured patient histories are represented as multi-channel trajectories:

$$X = \{(v_{i,j}, \tau_{i,j}, m_{i,j})\}_{i=1, j=1}^{C, T}, \quad (2)$$

where $v_{i,j}$, $\tau_{i,j}$, and $m_{i,j}$ denote the observed value, timestamp, and observation mask of variable i at time step j , respectively.

Textual context is constructed from chief complaint, diagnosis, past medical history, and historical treatment actions to provide disease-level semantic information. During training, future states in the target window are used only by the target branch for latent supervision and are never exposed to the prediction branch, preventing future information leakage. More details on scenario construction, sampling strategy, variable alignment, textual context, and leakage prevention are provided in Appendix A.

2.2 Training MedWorld-JEPA

We train the patient simulator as a latent clinical world model using a dual-branch Joint Embedding

Model	Hosp				ICU			
	RMSE ↓	MAE ↓	Range Agree ↑	Outcome ↑	RMSE ↓	MAE ↓	Range Agree ↑	Outcome ↑
Time Series Models								
MIRA	103.87	51.60	42.70%	—	90.51	59.79	—	—
Time-MoE-200M	132.45	42.58	39.67%	—	57.63	13.08	—	—
Large Language Models								
DeepSeek-V4-Pro	46.19	14.43	70.00%	78.08%	30.67	10.22	94.45%	89.06%
GPT-5.2	42.05	13.80	72.23%	82.50%	30.50	9.53	94.82%	83.62%
Claude Sonnet 4.5	30.22	7.58	80.17%	87.46%	60.43	9.73	83.27%	85.97%
Gemini-3.1-Pro	29.96	7.49	81.01%	93.28%	60.27	9.17	93.54%	89.65%
Kimi-K2.6	104.02	29.03	56.09%	57.32%	33.59	18.39	92.23%	79.65%
Qwen3-235B-A22B	261.54	42.11	66.90%	68.08%	13.91	9.39	90.81%	82.22%
MedWorld-JEPA	55.41	16.73	77.91%	91.72%	16.87	9.52	98.51%	94.16%

Table 2: Performance comparison under Hospital and ICU scenarios. Lower RMSE/MAE indicates better physiological prediction accuracy, while higher Range Agree and Outcome scores indicate better clinical consistency and outcome prediction performance.

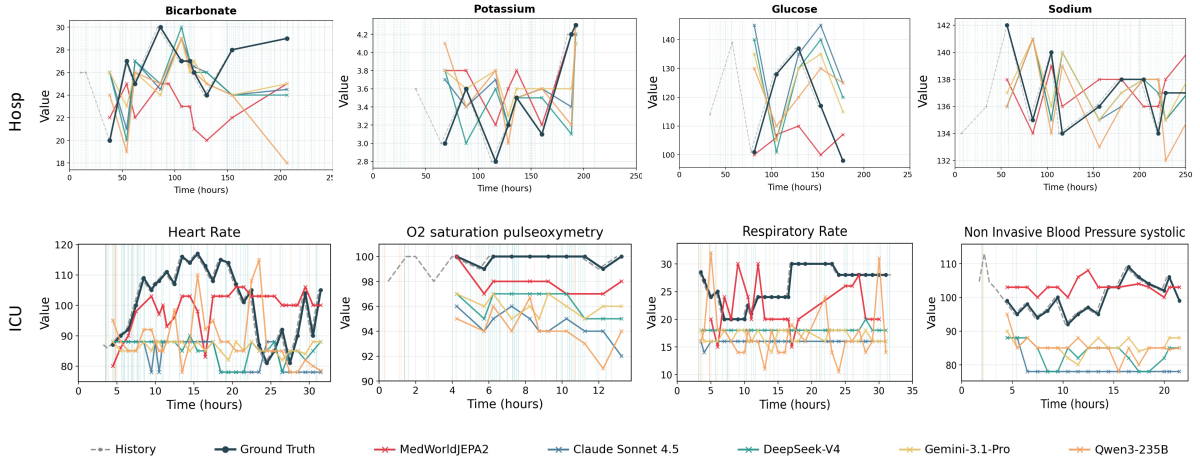


Figure 4: Comparison of physiological trajectory prediction under ICU and Hospital scenarios.

Predictive Architecture (JEPA) (Chen et al., 2025). We adopt this latent predictive framework because time-aware patient simulation requires modeling latent physiological evolution rather than directly generating sparse, noisy, and irregular clinical observations. Direct observation prediction may amplify noise and accumulate errors during long-horizon rollouts, while JEPA predicts future patient representations in a compact latent space, enabling the simulator to capture clinically meaningful state transitions driven by disease progression, treatment history, elapsed time, and clinical context. Observable clinical outputs are decoded from predicted latent states only when needed.

Model architecture. The patient simulator consists of a context encoder, a time- and action-conditioned predictor, a target encoder, and a clinical decoder. Given a historical patient trajectory

before time t , the context encoder encodes structured physiological time series, textual clinical context, clinical knowledge, and query information into a compact latent representation. Specifically, the structured time-series input $X_{\leq t}$ is encoded by a pretrained clinical temporal encoder based on MIRA (Li et al., 2026), while the textual clinical context c_t , disease-level knowledge d , and query text q are encoded by a frozen LLM encoder. The temporal and textual representations are integrated by a cross-modal fusion module:

$$h_t = F_\phi(E_{\text{ts}}(X_{\leq t}), E_{\text{text}}(c_t, d, q)), \quad (3)$$

where E_{ts} denotes the time-series encoder, E_{text} denotes the text encoder, and F_ϕ denotes the trainable fusion module.

To model clinical state transitions, we introduce a predictor that estimates the future latent patient

state conditioned on the historical latent context, elapsed time, and historical treatment actions:

$$\hat{z}_{t+\Delta} = P_\psi(h_t, e_\Delta, e_{a_{\leq t}}), \quad (4)$$

where P_ψ is the trainable predictor, e_Δ is the elapsed-time embedding, and $e_{a_{\leq t}}$ denotes the embedding of historical treatment actions. This time- and action-conditioned design allows the simulator to distinguish patient evolution under different temporal intervals and intervention histories, rather than assuming all transitions occur at fixed decision steps. The target branch encodes the real future window $(t, t + \Delta]$ into $z_{t+\Delta}^+$ only during training, providing latent supervision without exposing future information to the prediction branch.

Training objective. The simulator is trained to align the predicted future representation $\hat{z}_{t+\Delta}$ with the target future representation $z_{t+\Delta}^+$ in latent space. We use a symmetric contrastive objective, where each predicted–target pair from the same patient window is treated as a positive pair and other samples in the batch are treated as negatives. Let

$$s_{ij} = \text{sim}(\hat{z}_i, z_j^+)/\tau, \quad (5)$$

where τ is the temperature and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. The prediction loss is:

$$\mathcal{L}_{\text{pred}} = -\frac{1}{2B} \sum_{i=1}^B \left[\log \frac{\exp(s_{ii})}{\sum_{j=1}^B \exp(s_{ij})} + \log \frac{\exp(s_{ii})}{\sum_{j=1}^B \exp(s_{ji})} \right], \quad (6)$$

where B is the batch size. To avoid representation collapse, we add a uniformity regularization term:

$$\mathcal{L}_{\text{uni}} = \log \mathbb{E}_{i \neq j} \exp(-2\|\bar{z}_i - \bar{z}_j\|_2^2), \quad (7)$$

where $\bar{z}$ denotes ℓ_2 -normalized latent embeddings. The final latent dynamics learning objective is:

$$\mathcal{L}_{\text{latent}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{uni}} \mathcal{L}_{\text{uni}}. \quad (8)$$

Following a parameter-efficient training strategy, the MIRA backbone and base LLM encoder are frozen, while the fusion module, predictor, projection layers, and lightweight adapters are trainable. The target branch is updated using Exponential Moving Average (EMA):

$$\theta_{\text{target}} \leftarrow m\theta_{\text{target}} + (1 - m)\theta_{\text{online}}, \quad (9)$$

where m is the momentum coefficient.

Clinical decoding. After latent dynamics pre-training, a lightweight decoder maps predicted latent patient states into observable clinical outputs, including future laboratory values, vital signs, and clinical event descriptions. During environment interaction, the decoded observations are committed back to the patient trajectory, enabling multi-step simulation in the downstream asynchronous treatment environment. Additional implementation details are provided in Appendix B.

2.3 Construction of the MedACE

Based on the learned latent clinical world model, we construct an asynchronous treatment environment for long-horizon physician-agent interaction. For a patient trajectory, we denote the latent clinical state, observable measurements, physician action, and textual clinical context at time t as s_t , o_t , a_t , and c_t , respectively. Unlike conventional turn-based simulators where state transitions are triggered only by actions, i.e., $s_{k+1} = f(s_k, a_k)$, we model patient evolution as:

$$s_{t+\Delta t} = f_\theta(s_{\leq t}, a_{\leq t}, \Delta t, c_t, d), \quad (10)$$

where $s_{\leq t}$ and $a_{\leq t}$ denote historical patient trajectories and treatment actions, Δt denotes elapsed time, and d denotes disease-level knowledge. This formulation allows patient states to evolve under both natural disease progression and historical interventions.

At each interaction step, the physician agent performs a tool-based action, the environment advances the global clinical clock by Δt , and the simulator predicts future observations:

$$\hat{o}_{t+\Delta t} \sim p_\theta(o_{t+\Delta t} \mid o_{\leq t}, a_{\leq t}, c_t, d). \quad (11)$$

The simulator maintains the latent state, while decoded observations are exposed to the physician agent and appended to the observable patient trajectory. The elapsed time Δt can vary across steps according to the scenario and selected action, enabling asynchronous time advancement rather than a fixed round-based clock. The resulting interaction forms a closed loop of *observe* $\rightarrow$ *reason* $\rightarrow$ *act* $\rightarrow$ *simulate* $\rightarrow$ *observe*.

During training and evaluation, each episode corresponds to a patient admission trajectory and each step corresponds to a physician decision point. The action space covers diagnostic, monitoring, therapeutic, waiting, and terminal actions. To reduce

Model	Hosp				ICU			
	RMSE ↓	MAE ↓	Range Agree ↑	Outcome ↑	RMSE ↓	MAE ↓	Range Agree ↑	Outcome ↑
Qwen3-0.6B	7689.90	83.04	64.98%	73.55%	75.17	22.99	62.76%	80.14%
+ SFT	546.73	32.62	65.86%	78.54%	78.96	26.58	66.42%	86.75%
MIRA	103.87	51.60	42.70%	—	90.51	59.79	—	—
MedWorld-JEPA	55.41	16.73	77.91%	91.72%	16.87	9.52	98.51%	94.16%

Table 3: Ablation study under Hospital and ICU scenarios. Lower RMSE/MAE indicates better physiological prediction accuracy, while higher Range Agree and Outcome scores indicate better clinical consistency and downstream outcome prediction performance.

reward hacking from simulator-generated intermediate states, we use sparse rewards grounded in real clinical endpoints. The agent receives reward only when it executes the correct terminal action near the true terminal time:

$$r_t = \mathbb{I}[a_t = a^* \wedge |t - t^*| \leq \epsilon], \quad (12)$$

where a^* denotes the correct terminal action, t^* denotes the true terminal time, and ϵ denotes the allowable temporal tolerance window. This outcome-grounded design encourages agents to learn treatment selection, intervention timing, delayed treatment effects, and termination decisions. Additional environment details are provided in Appendix C.

3 Experiments

3.1 Main Results

Evaluation tasks and metrics. We evaluate the proposed framework on physiological state prediction and clinical outcome prediction. For physiological prediction, we assess numerical accuracy and clinical interpretability. In the Hosp scenario, MIMIC-IV provides normal-range annotations, so we compute Range Agree by measuring whether predicted values fall into the same clinical range as the ground truth, indicating whether the model correctly identifies normal or abnormal physiological states. In the ICU scenario, where explicit normal ranges are unavailable but warning labels indicate critical thresholds, we prompt generative models to output binary warning predictions and use them to compute ICU Range Agree. For outcome prediction, patients who die during hospitalization or are discharged to hospice care and die within six months are defined as treatment failures, while the others are considered recovered. The model predicts whether the patient should be discharged at the current time, whether discharge will occur within the next 4h or 12h, and whether no significant event will occur within the next 12h. We

use prediction accuracy as the outcome evaluation metric.

Baselines. We compare two categories of models. The first includes traditional time-series forecasting models, namely the specialized medical temporal model MIRA (Li et al., 2026) and the general time-series model Time-MoE-200M (Shi et al., 2025). The second includes state-of-the-art LLMs, including DeepSeek-V4-Pro (DeepSeek-AI, 2026), GPT-5.2 (Singh et al., 2025), Claude Sonnet 4.5 (Anthropic, 2025), Gemini-3.1-Pro (Google AI for Developers, 2026), Kimi-K2.6 (Moonshot AI, 2026), and Qwen3-235B-A22B (Yang et al., 2025). Compared with conventional time-series models, LLMs can leverage richer textual context, including treatment actions, patient history.

Quantitative results. Table 2 reports the experimental results. Traditional time-series models generally show higher prediction error and lower clinical consistency, as they rely only on structured temporal inputs and cannot effectively use treatment actions or textual clinical knowledge. This limitation is especially evident in the Hosp scenario, where sparse temporal signals make MIRA and Time-MoE substantially worse than LLM-based methods in RMSE and MAE. LLM baselines achieve stronger overall performance, suggesting the importance of textual semantics and disease knowledge for clinical state prediction. However, without explicit temporal dynamics modeling, they remain unstable in the ICU scenario, particularly for high-frequency physiological fluctuations and abnormal transitions. Our model achieves the best overall ICU performance, significantly outperforming all baselines in RMSE, Range Agree, and outcome prediction accuracy, showing that temporal encoders, asynchronous evolution modeling, and disease knowledge help capture long-term clinical dynamics and delayed treatment effects.

Environment	Time-aware	Natural Progression	Action-conditioned	RMSE ↓	MAE ↓	Range Agree ↑	Outcome Acc. ↑
Hosp							
Action-based	×	×	✓	714.37	55.56	60.82%	63.91%
Time-based (MIRA)	✓	✓	×	103.87	51.60	42.70%	–
w/o Timestamp	×	✓	✓	593.20	47.60	68.09%	72.64%
Ours	✓	✓	✓	55.41	16.73	77.91%	91.72%
ICU							
Action-based	×	×	✓	106.43	32.06	62.51%	51.34%
Time-based (MIRA)	✓	✓	×	90.51	59.79	–	–
w/o Timestamp	×	✓	✓	91.82	28.56	67.12%	81.56%
Ours	✓	✓	✓	16.87	9.52	98.51%	94.16%

Table 4: Ablation study on environment modeling strategies under Hospital and ICU scenarios. Time-aware indicates whether explicit temporal evolution is modeled; Natural Progression denotes whether disease progression is modeled independently of physician actions; Action-conditioned denotes whether patient transitions depend on treatment actions. Our full model consistently outperforms all ablations, highlighting the importance of jointly modeling temporal dynamics, natural disease progression, and intervention effects.

Qualitative trajectory analysis. Figure 4 further provides a detailed comparison of predicted physiological trajectories. MedWorldJEPA better preserves the temporal trend of ground-truth clinical indicators, especially when patient states undergo abrupt increases or decreases. In contrast, LLM-based baselines often exhibit delayed responses: their predictions tend to follow previous historical trends rather than anticipate the next physiological transition. This suggests that general-purpose LLMs are largely imitating past trajectory patterns from the observed context, rather than truly modeling future patient dynamics. Such lagging behavior is particularly problematic in ICU scenarios, where physiological variables change rapidly and delayed prediction may lead to incorrect clinical interpretation. These qualitative results further support our quantitative findings that explicit temporal encoding and asynchronous state-transition modeling are essential for accurate patient simulation.

3.2 Ablation Study

Table 3 presents the ablation study under both Hospital and ICU scenarios. Starting from the base LLM (Qwen3-0.6B), supervised fine-tuning (+SFT) significantly improves both physiological prediction and outcome prediction performance, demonstrating the effectiveness of leveraging textual clinical context and disease knowledge. However, since the +SFT setting does not include the temporal encoder and relies solely on textual modeling, its performance remains limited in scenarios with complex physiological dynamics, particularly in the ICU setting. In contrast, MIRA, which only utilizes structured temporal modeling without tex-

tual semantics, achieves relatively lower RMSE in some settings but suffers from poor clinical consistency and lacks effective outcome prediction ability, indicating that time-series information alone is insufficient for complex clinical transitions.

Our full model achieves the best overall performance across both Hospital and ICU scenarios. By jointly integrating temporal encoding, asynchronous temporal evolution modeling, and textual clinical semantics, the proposed framework significantly improves RMSE, Range Agree, and outcome prediction accuracy. The improvement is especially pronounced in the ICU setting, demonstrating that combining structured physiological dynamics with disease-level semantic knowledge is critical for modeling long-term patient trajectories and delayed treatment effects.

3.3 Doctor Agent Evaluation

We evaluate different models as doctor agents in the constructed virtual treatment environment. As shown in Table 5, general-purpose LLMs without environment training commonly exhibit severe *early discharge* behavior. In both Hosp and ICU scenarios, most models terminate treatment after only a few interaction steps, indicating that although they possess medical knowledge, they have not learned to continuously observe patient states, wait for delayed feedback, or make long-horizon decisions in a dynamic environment. Among these models, Claude Sonnet 4.5 achieves the highest Hosp success rate of 26.00% with 21.56 average steps, showing relatively stronger sustained interaction and treatment planning ability. However, its Failure / Dead rate reaches 9.33%, suggesting that a

Model	Success	Failure	Early Discharge	Steps
Hosp				
Claude Sonnet 4.5	26.00%	9.33%	64.67%	21.56
Gemini-3.1-Pro	13.33%	3.33%	83.33%	6.63
GPT-5.2	10.67%	1.33%	88.00%	5.96
DeepSeek-V4-Pro	9.33%	0.67%	90.00%	5.49
Qwen3-235B-A22B	7.33%	9.33%	82.67%	3.32
Qwen3-8B	6.67%	24.00%	69.33%	5.48
+SFT	15.42%	18.54%	66.04%	9.25
+RL	35.68%	23.65%	40.67%	25.76
ICU				
Claude Sonnet 4.5	5.92%	1.32%	92.76%	9.45
Gemini-3.1-Pro	3.29%	0.66%	96.05%	5.78
GPT-5.2	2.63%	0.66%	96.71%	5.26
DeepSeek-V4-Pro	1.32%	0.00%	98.68%	4.11
Qwen3-235B-A22B	1.32%	4.61%	94.08%	3.70
Qwen3-8B	1.97%	17.76%	80.26%	3.92
+SFT	3.95%	19.74%	76.31%	15.58
+RL	15.79%	36.18%	48.03%	29.62

Table 5: Evaluation results of LLM-based physician agents in the constructed virtual treatment environment.

longer interaction horizon alone does not guarantee better treatment quality.

For trainable models, supervised fine-tuning (SFT) helps the model acquire initial long-horizon reasoning and interaction behaviors by imitating more reasonable clinical patterns and reducing premature termination. The main improvement comes from reinforcement learning (RL). In the Hosp scenario, RL increases the Success rate to 35.68%, outperforming all general-purpose LLM baselines, and reduces Early Discharge to 40.67%. In the ICU scenario, RL improves the Success rate from 3.95% to 15.79% and reduces Early Discharge from 76.31% to 48.03%. The Avg. Step also increases to 25.76 and 29.62 in Hosp and ICU, respectively, indicating that RL substantially enhances long-horizon treatment decision-making by discouraging premature discharge and encouraging extended interaction.

However, RL introduces risks. As the model reduces early termination and performs longer, more complex interventions, the Failure / Dead rate increases from 18.54% to 23.65% in Hosp and from 19.74% to 36.18% in ICU. This suggests that although the RL strategy mitigates early discharge and improves treatment success, it may also lead to prolonged or insufficiently optimized interventions for high-risk patients, increasing treatment failure. Overall, these results show that the proposed environment can drive models from short-sighted termination toward long-horizon interactive decision-making, while future work should refine reward design to better balance premature-discharge reduction, risk control, and failure prevention.

4 Related Work

Recent research on LLM agents has shifted from static instruction-following datasets to interactive training and evaluation environments, where agents learn through multi-turn action-observation trajectories, tool use, and verifiable feedback (Tu et al., 2026; Song et al., 2026; Xiao et al., 2026). This trend highlights the need for scalable, executable, and realistic environments for SFT/RL training. Recent studies further note that many benchmarks still assume a sequential protocol in which the environment is paused while the agent acts, failing to capture asynchronous real-world dynamics where states evolve independently, feedback may be delayed, and new events may emerge (Froger et al., 2025). A similar shift is emerging in medical agents, where evaluation is moving from static medical question answering to interactive clinical environments that reflect real diagnostic and treatment workflows (Schmidgall et al., 2024; Li et al., 2024; Qiu et al., 2025; Choudhary et al., 2024; Wu et al., 2026; Zhao et al., 2026; Zhu et al., 2026). These environments increasingly emphasize multi-turn information gathering, examination selection, tool/API interaction, patient state transitions, and closed-loop evaluation. Following clinical environment simulator frameworks that suggest jointly modeling disease progression, intervention responses, temporal delays, and hospital-level resources (Luo et al., 2026), our work focuses on treatment-oriented asynchronous patient dynamics, where patient states evolve over time and treatment effects may appear only after a delay.

5 Conclusions

We propose a time-aware asynchronous clinical environment, MedACE, for treatment-oriented medical agents. By modeling physiological trajectories, textual clinical context, elapsed time, and treatment actions, it captures natural disease progression and delayed intervention effects beyond turn-based simulators. Built on a latent patient simulator, MedWorld-JEPA, it supports long-horizon, closed-loop physician-agent interaction with outcome-grounded evaluation. Experiments in Hospital and ICU scenarios demonstrate effectiveness in physiological forecasting, clinical consistency, and agent evaluation, while revealing limitations of current LLM-based agents in long-horizon treatment decision-making and the need for realistic dynamic environments and safer reward designs.

Limitations

This work mainly focuses on building a time-aware asynchronous clinical environment for long-horizon medical treatment agents. While we also perform preliminary reinforcement learning for doctor-agent training, the current setting only adopts GRPO with outcome-based rewards. This sparse reward design may not fully capture fine-grained clinical requirements such as intervention timing, risk control, and safe treatment planning.

Experimental results show that RL reduces early discharge and improves treatment success, but also leads to instability, including an increased Failure / Dead rate after training. This indicates that doctor-agent training in dynamic clinical environments requires more carefully designed reward functions and RL algorithms. Future work will further explore safety-aware reward modeling, denser clinical feedback, and more stable optimization strategies for training doctor agents.

Ethical Considerations

This work is intended only for research on simulated medical agent evaluation and should not be used for real-world diagnosis, treatment recommendation, triage, or autonomous clinical deployment. Any clinical application would require rigorous prospective validation, regulatory review, and supervision by qualified healthcare professionals. The data used in this study are derived from de-identified EHR records, and researchers must follow the corresponding data use agreements and privacy requirements. We do not attempt to re-identify patients or release individual-level records. A key ethical risk is that simulated doctor agents may produce unsafe, biased, or overconfident treatment decisions. Since the simulator approximates clinical dynamics, its outputs should be interpreted as experimental simulation signals rather than medical advice.

Acknowledgments

This work is supported by the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (Grant no. 2025C01128), the National Natural Science Foundation of China (62476241), and the ZJU-Angelalign R&D Research Center.

References

- Anthropic. 2025. Introducing claude sonnet 4.5. <https://www.anthropic.com/news/claude-sonnet-4-5>. Accessed: 2026-05-26.
- Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Yejin Bang, Allen Bolourchi, Yann LeCun, and Pascale Fung. 2025. Vl-jepa: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*.
- Kartik Choudhary, Dhawal Gupta, and Philip S Thomas. 2024. Icu-sepsis: A benchmark mdp built from real medical data. *arXiv preprint arXiv:2406.05646*.
- DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.
- Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, and 1 others. 2025. Are: Scaling up agent environments and evaluations. *arXiv preprint arXiv:2509.17158*.
- Google AI for Developers. 2026. Gemini 3 developer guide. <https://ai.google.dev/gemini-api/docs/gemini-3>. Last updated: 2026-05-07; accessed: 2026-05-26.
- Songtao Jiang, Yuan Wang, Sibao Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, and 1 others. 2025a. Hulumed: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*.
- Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. 2025b. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *Nejm Ai*, 2(9):A1dbp2500144.
- Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, and 1 others. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature medicine*, 31(1):77–86.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

- Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and 1 others. 2026. Mira: Medical time series foundation model for real-world health data. *Advances in Neural Information Processing Systems*, 38:98657–98685.
- Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and 1 others. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.
- Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Luyang Luo, Sung Eun Kim, Xiaoman Zhang, Julius M Kernbach, Roshan Kenia, Julian N Acosta, Larry A Nathanson, Adrian D Haimovich, Adam Rodman, Ethan Goh, and 1 others. 2026. A clinical environment simulator for dynamic ai evaluation. *Nature medicine*, pages 1–8.
- Moonshot AI. 2026. Kimi k2.6: Advancing open-source coding. <https://www.kimi.com/blog/kimi-k2-6>. Accessed: 2026-05-26.
- Pengcheng Qiu, Chaoyi Wu, Junwei Liu, Qiaoyu Zheng, Yusheng Liao, Haowen Wang, Yun Yue, Qianrui Fan, Shuai Zhen, Jian Wang, and 1 others. 2025. Evolving diagnostic agents in a virtual clinical environment. *arXiv preprint arXiv:2510.24654*.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cian Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. 2024. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22315–22339.
- Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. 2025. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, volume 2025, pages 34635–34667.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.
- Xiaoshuai Song, Haofei Chang, Guanting Dong, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. 2026. Envs-caler: Scaling tool-interactive environments for llm agent via programmatic synthesis. *arXiv preprint arXiv:2601.05808*.
- Dunwei Tu, Hongyan Hao, Hansi Yang, Yihao Chen, Yi-Kai Zhang, Zhikang Xia, Yu Yang, Yueqing Sun, Xingchen Liu, Furao Shen, and 1 others. 2026. Scaleenv: Scaling environment synthesis from scratch for generalist interactive tool-use agent training. *arXiv preprint arXiv:2602.06820*.
- Minghao Wu, Yuting Yan, Zhenyang Cai, Ke Ji, Chuansen Fang, Ziyang Sheng, Xidong Wang, Rongsheng Wang, Hejia Zhang, Shuang Li, and 1 others. 2026. Agentifying patient dynamics within llms through interacting with clinical world model. *arXiv preprint arXiv:2605.14723*.
- Zikai Xiao, Jianhong Tu, Chuhang Zou, Yuxin Zuo, Zhi Li, Peng Wang, Bowen Yu, Fei Huang, Junyang Lin, and Zuozhu Liu. 2026. Webworld: A large-scale world model for web agent training. *arXiv preprint arXiv:2602.14721*.
- Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Zifeng Wang, Xiangru Tang, Hang Wu, May D Wang, Peifeng Ruan, Donghan Yang, and 1 others. 2025. Medagentgym: A scalable agentic training environment for code-centric reasoning in biomedical data science. *arXiv preprint arXiv:2506.04405*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, and 1 others. 2023. Huatuoogpt, towards taming language model to be a doctor. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 10859–10885.
- Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, and 1 others. 2026. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, pages 1–10.
- Yinghao Zhu, Ziyi He, Haoran Hu, Xiaochen Zheng, Xichen Zhang, Wang Wang, Junyi Gao, Liantao Ma, and Lequan Yu. 2026. Medagentboard: Benchmarking multi-agent collaboration with conventional methods for diverse medical tasks. *Advances in Neural Information Processing Systems*, 38.

Use of AI Assistance

We used an AI-based writing assistant to support language polishing, grammar correction, and clarity improvement during manuscript preparation. The assistant was also used to help identify potential inconsistencies, formatting issues, and minor detail errors in the manuscript. All technical ideas, experimental design, data analysis, results interpretation, and final conclusions were developed, verified, and approved by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the submitted work.

A Time-Aware Trajectory Construction Details

Scenario Construction We construct two clinical scenarios according to the care setting of each patient trajectory. The Hospital scenario is designed for general hospitalized patients, where clinical observations are relatively sparse and patient state evolution usually reflects longer-term disease progression. The ICU scenario is designed for critically ill patients, where physiological measurements and treatment actions are recorded more frequently, requiring higher temporal resolution to capture rapid deterioration and short-term physiological fluctuations.

The two scenarios therefore use different temporal resolutions. This design allows the patient simulator to learn both long-term inpatient progression and high-frequency critical care dynamics within a unified time-aware modeling framework.

Sliding Window Sampling For each patient trajectory, we generate supervised samples using a sliding temporal window strategy. Given a reference time point t , the historical context window $[t - H, t]$ is used as model input, while the future target window $(t, t + \Delta]$ provides supervision for future state prediction. Here, H denotes the history length and Δ denotes the prediction horizon.

In the Hospital scenario, we set $H = 48h$ and $\Delta = 4h$, enabling the model to capture long-range clinical progression. In the ICU scenario, we set $H = 4h$ and $\Delta = 0.5h$, allowing the model to focus on high-frequency physiological changes. Historical observations, treatment actions, and textual clinical context are temporally aligned within the same window.

Clinical Variables and Action Alignment Structured time-series variables include laboratory val-

ues and vital signs, such as creatinine, glucose, heart rate, blood pressure, oxygen saturation, and other clinically relevant measurements. Since real-world clinical records are irregularly sampled and contain missing values, each variable is represented by its observed value, timestamp, and observation mask:

$$X = \{(v_{i,j}, \tau_{i,j}, m_{i,j})\}_{i=1, j=1}^{C,T}.$$

This representation preserves temporal information and enables the model to handle irregular sampling, missing observations, and heterogeneous recording frequencies.

Historical treatment actions are aligned with the observation window. These actions include medications, procedures, laboratory orders, monitoring events, and other intervention-related records. Depending on the input modality, treatment actions are incorporated into either the structured action history or the textual clinical context, allowing the simulator to condition future state prediction on both natural disease progression and intervention history.

Textual Clinical Context In addition to structured time-series data, we construct textual clinical context to provide semantic information that may not be fully captured by numerical observations. The context integrates chief complaint, diagnosis, past medical history, and historical treatment actions, enabling the model to incorporate disease-level priors and long-range clinical semantics.

We also construct query text to specify the prediction target. For example, the query may ask the model to predict a future laboratory value or clinical event within a specific horizon:

“Predict future potassium level in the next 4 hours.”

This query-conditioned design allows the simulator to focus on specific future clinical variables or events while sharing the same historical patient context.

Target Construction and Leakage Prevention The real patient states within the future target window $(t, t + \Delta]$ are organized as target representations for latent supervision. These targets may include future physiological measurements, clinical events, and target text constructed from the future window.

Importantly, future information is used only by the target branch during training. The prediction

branch only receives historical observations, historical treatment actions, textual clinical context, disease-level knowledge, and elapsed-time information. During inference and environment interaction, the target branch is removed, and future patient states are predicted solely from historical information. This design prevents future information leakage while enabling JEPA-style latent future state prediction.

Patient-Level Data Split Because multiple sliding-window samples can be generated from the same patient trajectory, we split the dataset at the patient level. This ensures that samples from the same patient do not appear across training, validation, and test sets, reducing trajectory leakage and providing a more reliable evaluation of generalization to unseen patients.

B Latent Clinical World Model Details

Architecture Details The latent clinical world model follows a dual-branch JEPA-style architecture. The prediction branch, also referred to as the online branch, receives only historical information, including structured physiological time series, textual clinical context, historical treatment actions, disease-level knowledge, query text, and elapsed-time condition. The target branch receives the real future window and is used only during training to provide latent supervision.

The context encoder contains two modality-specific encoders. The time-series encoder is initialized from the pretrained clinical temporal model MIRA, which is designed for irregularly sampled physiological trajectories. The text encoder is implemented with a frozen LLM encoder to provide clinical semantic representations from textual context, disease knowledge, and query text. The output representations from the two encoders are fused by a trainable cross-modal fusion module. In our implementation, only the fusion module, predictor, projection layers, lightweight adapters, decoder, and other task-specific heads are updated during training, while the MIRA backbone and base LLM encoder remain frozen.

Time and Action Conditioning To support time-aware patient state transition, the predictor is explicitly conditioned on elapsed time and historical treatment actions. The elapsed-time embedding e_Δ represents the prediction horizon between the current reference time t and the future target time $t + \Delta$.

The treatment-action embedding $e_{a \leq t}$ summarizes historical interventions, including medications, procedures, laboratory orders, monitoring events, and other clinically relevant actions.

This design allows the simulator to model different future trajectories under different elapsed-time intervals and intervention histories. For example, the same patient state may evolve differently after a short waiting period, a delayed treatment, or an immediate intervention. Therefore, conditioning on both elapsed time and action history is essential for modeling natural disease progression and delayed treatment effects.

Target Branch and Leakage Prevention The target branch encodes the real patient state within the future target window $(t, t + \Delta]$. Its input may include future physiological measurements, clinical events, and target text constructed from the future window. The target representation $z_{t+\Delta}^+$ is used as latent supervision for the prediction branch.

Importantly, the target branch is used only during training. Future observations and target text are never provided to the prediction branch. During inference and environment interaction, the target branch is removed, and the simulator predicts future latent states solely from historical observations, historical treatment actions, textual clinical context, disease-level knowledge, and elapsed-time information. This prevents future information leakage while enabling latent future state prediction.

Training Details The simulator is trained with a latent prediction objective and a representation regularization objective. The symmetric contrastive loss aligns predicted future representations with target future representations, while using other samples in the batch as negatives. The uniformity regularization term encourages latent embeddings to remain well distributed and reduces the risk of representation collapse.

The final latent training objective is:

$$\mathcal{L}_{\text{latent}} = \mathcal{L}_{\text{pred}} + \lambda_{\text{uni}} \mathcal{L}_{\text{uni}}. \quad (13)$$

The target branch is updated using EMA rather than direct gradient updates, which stabilizes the latent target space and improves long-horizon dynamics modeling.

Clinical Decoder After latent dynamics pretraining, we train a lightweight clinical decoder to map predicted latent patient states back to observable

outputs. For continuous variables such as laboratory values and vital signs, the decoder is optimized with regression losses. For discrete clinical events or outcome-related descriptions, the decoder can be optimized with classification or language modeling losses.

During environment interaction, the predicted latent state is decoded into observable clinical outputs. These outputs are exposed to the physician agent and appended to the observable patient trajectory, allowing the downstream treatment environment to perform multi-step closed-loop simulation.

Implementation Notes The latent clinical world model is designed to decouple patient dynamics modeling from observation generation. The prediction branch learns the latent transition of patient states, while the decoder produces observable clinical feedback only when required by the treatment environment. This design reduces the burden of directly generating heterogeneous clinical records at every step and improves stability during long-horizon rollouts.

C Asynchronous Treatment Environment Details

Environment State and Observation The asynchronous treatment environment is built upon the learned latent clinical world model. Internally, the simulator maintains a latent patient state s_t , which summarizes the patient’s physiological condition, disease progression, treatment history, and clinical context up to time t . The physician agent does not directly observe this latent state. Instead, it receives decoded observable clinical feedback o_t , including laboratory values, vital signs, clinical events, and textual descriptions generated from the simulator.

At each step, the environment appends the newly predicted observation $\hat{o}_{t+\Delta t}$ to the observable patient trajectory. This allows the agent to interact with a continuously evolving patient history rather than a static case description.

Action Space The physician agent interacts with the environment through tool-based actions. We organize the action space into five categories:

- **Diagnostic actions:** ordering laboratory tests, imaging examinations, or other diagnostic procedures.
- **Monitoring actions:** checking vital signs, reviewing laboratory results, or monitoring changes in patient condition.

- **Therapeutic actions:** prescribing medications, performing procedures, or adjusting treatment plans.
- **Waiting actions:** choosing no immediate intervention and allowing time to advance.
- **Terminal actions:** discharging the patient or terminating treatment when a final clinical endpoint is reached.

This action design allows the environment to evaluate not only which treatment action is selected, but also whether the agent chooses appropriate diagnostic, monitoring, waiting, or terminal decisions over a long-horizon trajectory.

Asynchronous Time Advancement Unlike turn-based simulators that update patient states only after discrete action rounds, our environment maintains a global clinical clock. After the physician agent executes an action, the environment advances time by an elapsed interval Δt . The value of Δt can depend on both the clinical scenario and the selected action.

For example, the Hospital scenario uses a coarser temporal resolution to reflect slower inpatient disease progression, while the ICU scenario uses a finer temporal resolution to capture rapid physiological changes. In addition, different actions may imply different time costs: waiting directly advances the clock, diagnostic actions may require time before results become available, and therapeutic actions may have delayed physiological effects. This design supports asynchronous patient evolution and allows the simulator to model both natural disease progression and delayed treatment responses.

Closed-Loop Simulation Procedure At each interaction step, the environment follows the closed-loop procedure below:

1. The physician agent observes the current patient information $o_{\leq t}$ and textual clinical context c_t .
2. The agent reasons over the patient history and selects a tool-based action a_t .
3. The environment advances the global clock by Δt according to the scenario and action.
4. The latent clinical world model predicts the future patient state conditioned on historical

observations, historical actions, elapsed time, clinical context, and disease-level knowledge.

5. The predicted latent state is decoded into observable clinical outputs $\hat{o}_{t+\Delta t}$.
6. The decoded observations are appended to the patient trajectory and exposed to the agent at the next step.

This process forms a closed loop of *observe* $\rightarrow$ *reason* $\rightarrow$ *act* $\rightarrow$ *simulate* $\rightarrow$ *observe*, enabling long-horizon physician-agent interaction with asynchronously evolving patient states.

Outcome-Grounded Reward Construction To reduce reward hacking caused by simulator-generated intermediate feedback, we use sparse rewards grounded in real clinical endpoints rather than dense rewards derived from simulated states. Each episode is associated with a real patient outcome. If the patient is successfully discharged, the correct terminal action is discharge; if the patient reaches treatment failure or death, the correct terminal action is finish_treatment. The reward is defined as:

$$r_t = \mathbb{I}[a_t = a^* \wedge |t - t^*| \leq \epsilon], \quad (14)$$

where a^* denotes the correct terminal action, t^* denotes the true terminal time, and ϵ is the allowable temporal tolerance window.

This reward design encourages the agent to identify both the correct terminal decision and the appropriate timing. Intermediate treatment behaviors are encouraged indirectly through their influence on future simulated patient states and terminal timing. As a result, the environment evaluates whether an agent can balance treatment selection, delayed treatment effects, continuous monitoring, and termination decisions over long clinical horizons.