Learning Dynamic Representations and Policies from Multimodal Clinical Time-Series with Informative Missingness

Zihan Liang* **Ziwen Pan*** **Ruoxuan Xiong**
Emory University, Atlanta, USA
{zihan.liang, ziwen.pan, ruoxuan.xiong}@emory.edu

Abstract

Multimodal clinical records contain structured measurements and clinical notes recorded over time, offering rich temporal information about the evolution of patient health. Yet these observations are sparse, and whether they are recorded depends on the patient's latent condition. Observation patterns also differ across modalities, as structured measurements and clinical notes arise under distinct recording processes. While prior work has developed methods that accommodate missingness in clinical time series, how to extract and use the information carried by the observation process itself remains underexplored. We therefore propose a patient representation learning framework for multimodal clinical time series that explicitly leverages informative missingness. The framework combines (1) a multimodal encoder that captures signals from structured and textual data together with their observation patterns, (2) a Bayesian filtering module that updates a latent patient state over time from observed multimodal signals, and (3) downstream modules for offline treatment policy learning and patient outcome prediction based on the learned patient state. We evaluate the framework on ICU sepsis cohorts from MIMIC-III, MIMIC-IV, and eICU. It improves both offline treatment policy learning and adverse outcome prediction, achieving FQE 0.679 versus 0.528 for clinician behavior and AUROC 0.886 for post-72-hour mortality prediction on MIMIC-III.

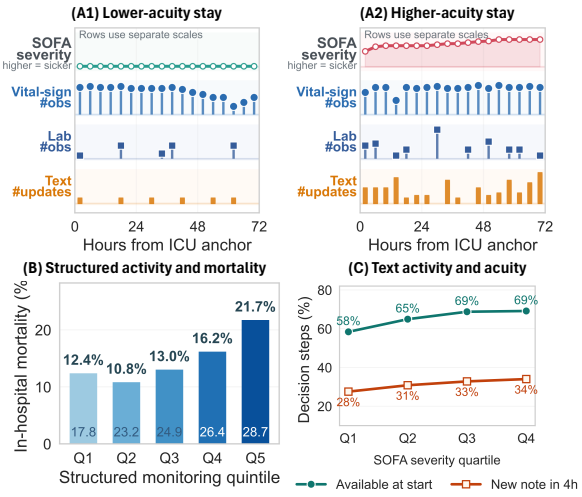

Figure 1: **Text and structured observations exhibit temporal MNAR patterns in ICU care.** **(A)** Two representative MIMIC-III ICU stays illustrate how acuity and observation processes co-evolve. The top row shows SOFA severity (higher = sicker); the lower rows show counts of vital-sign observations, laboratory observations, and text updates within each 4-hour bin. Compared with a lower-acuity stay (A1), a higher-acuity stay (A2) shows rising SOFA, denser laboratory measurements, and burstier documentation. **(B)** Across ICU stays, greater structured monitoring intensity is associated with higher in-hospital mortality. **(C)** Clinical text follows the same endogenous pattern: higher-acuity decision steps are more likely to already have text available and to receive new text within the next 4 hours.

1 Introduction

Electronic health records are multimodal, consisting of structured data, such as vital signs and laboratory measurements, as well as clinical texts, such as notes and reports. These data are recorded longitudinally and encode rich temporal information about how patient health evolves over time. This makes it possible to learn dynamic patient representations that support both outcome prediction and sequential clinical decision-making. However, two key features exist in clinical observations that complicate how such representations should be learned.

First, they are often sparse and irregular, and their observation process depends on both clinician decisions and the patient's underlying health state. Figure 1(A) illustrates this pattern using two representative ICU trajectories from MIMIC-III (Johnson et al., 2016): compared with a lower-acuity tra-

*Equal contribution. Code and reproducibility materials are available at https://github.com/CausalMLResearch/OPL-MT-MNAR under the MIT License.

Findings of the Association for Computational Linguistics: ACL 2026, pages 26363–26392
July 2-7, 2026

jectory, a higher-acuity trajectory exhibits denser laboratory monitoring and more frequent text updates. The same pattern also appears at the cohort level. Using MIMIC-IV (Johnson et al., 2023), Figure 1(B) shows that greater structured monitoring intensity is associated with higher in-hospital mortality. Figure 1(C) shows a parallel relationship for text: higher-acuity patients are more likely to have text available at a given decision step and to receive new text in the following step.

Second, observation patterns differ systematically across modalities because different types of clinical data are generated through different mechanisms. Vital signs are often collected more routinely, laboratory tests need to be ordered, and text updates depend even more directly on clinician documentation behavior. This contrast is visible in Figure 1(A): even within the same patient trajectory, the temporal availability of structured measurements and text updates evolves differently.

Taken together, these patterns suggest that observation processes are informative about patient state but should be used carefully, as their meanings differ across modalities. For structured clinical time series, prior work has proposed methods that incorporate informative missingness through masks and time gaps (Che et al., 2018). In the multimodal setting, Liang et al. (2025) also explicitly models informative missingness, but without temporal dynamics. However, it remains open how to leverage informative missingness over time across modalities to learn patient representations.

We propose OPL-MT-MNAR (Off-Policy Learning under Multimodal Observations with Temporal Missing-Not-At-Random Patterns), a framework that explicitly leverages informative missingness in multimodal clinical records. Our approach has two stages. In the first stage, we learn a dynamic latent representation from the multimodal observations available up to the current time. In the second stage, we use the learned patient representation for outcome prediction and offline policy learning.

The first stage is motivated by *Bayesian filtering* and consists of two components. The first is a multimodal encoder that learns a unified representation from the data observed so far. For structured data, we construct the embedding using an extension of Gated Recurrent Units-Decay (GRU-D) (Che et al., 2018) together with additional missingness-aware features (time since last observation, cumulative observation counts, missing rates, and windowed observation frequency). For clinical text, we introduce a temporal documentation factor that is updated at each time step and summarizes the observation pattern of the text observed so far. This factor is then used both to refine the text embedding and to guide the fusion of text and structured-data embeddings into a unified representation.

The second component models *patient dynamics* through a latent belief state learned via variational inference. This belief state captures underlying health dynamics together with the cumulative effects of past treatment actions. It is then combined with the unified representation from the multimodal encoder to form a posterior patient state for downstream tasks. We *show theoretically* that without such a belief state, the multimodal encoder alone may fail to preserve sufficient information about treatment history for sequential decision-making.

In the second stage, we use the posterior patient state for multiple downstream tasks. One task is offline treatment policy optimization using expectile regression (Kostrikov et al., 2022), which accommodates delayed rewards. The other is outcome prediction. Jointly learning these tasks allows the shared patient representation to benefit from positive transfer across tasks.

We evaluate our framework on MIMIC-III, MIMIC-IV, and eICU (Pollard et al., 2018), which together cover complementary text-observation regimes and cross-institutional generalization. Our empirical focus is ICU sepsis care with 4-hour decision steps over the first 72 ICU hours. Across these settings, the learned state supports both clinically meaningful prediction and offline treatment optimization, with the largest gains appearing in high-acuity settings where observation processes are most informative. Concretely, the OPL-MT-MNAR policy achieves FQE 0.679 on MIMIC-III, 0.634 on MIMIC-IV, and 0.604 on eICU, compared with clinician behavior at 0.528, 0.521, and 0.534, respectively. Using the same learned representation, we also achieve AUROC 0.886 for post-72-hour mortality prediction on MIMIC-III, with the clearest policy improvements appearing in the highest-acuity subgroup.

2 Problem Formulation

We consider a finite-horizon partially observable Markov decision process (POMDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, P, \Omega, R, \gamma, H)$ (Kaelbling et al., 1998; Hauskrecht and Fraser, 2000), where $\mathcal{S}$ is the latent state space, $\mathcal{A}$ is the discrete action space, $\mathcal{O}$ is

the observation space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $\Omega : \mathcal{S} \to \Delta(\mathcal{O})$ is the observation function, R is the reward function, $\gamma \in [0, 1)$ is the discount factor, and H is the horizon. The true patient state $s_h \in \mathcal{S}$ at decision step h is not directly observed.

At decision step h, the agent *partially observes* the following information

$$o_h = (\boldsymbol{y}_h^{\mathsf{s}}, \boldsymbol{m}_h^{\mathsf{s}}, \boldsymbol{y}_h^{\mathsf{t}}, \boldsymbol{m}_h^{\mathsf{t}}) \in \mathcal{O}.$$

For structured data, $\boldsymbol{y}_h^{\mathsf{s}} \in \mathbb{R}^{|\mathcal{T}_h| \times D}$ denotes the measurement values recorded at observation times $\mathcal{T}_h$ within decision step h, and $\boldsymbol{m}_h^{\mathsf{s}} \in \{0, 1\}^{|\mathcal{T}_h| \times D}$ indicates whether each entry is observed. For text data, $\boldsymbol{y}_h^{\mathsf{t}} = \{y_h^{\mathsf{t},j}\}_{j \in \mathcal{M}^{\mathsf{t}}}$ denotes the collection of raw text observations across modalities, and $\boldsymbol{m}_h^{\mathsf{t}} = [m_h^{\mathsf{t},j}]_{j \in \mathcal{M}^{\mathsf{t}}}$ is the corresponding binary indicator vector, where $m_h^{\mathsf{t},j}$ records whether text modality j is observed at step h. These raw text observations are encoded into a step-level text embedding $e_h^{\mathsf{t}} \in \mathbb{R}^{d_e}$ before entering the multimodal fusion module.

In addition, static patient features x (e.g., age and gender) are available throughout the trajectory. We let $I_h = \{x, o_1, \cdots, o_h\}$ be the information set accumulated up to step h, comprising static features together with structured and textual information collected dynamically.

Episodes may terminate at step $h^* \leq H$ if the patient dies, is discharged, or reaches the end of the observation window. The logged episode ends at h^*, so no observations, actions, or rewards are defined for $h > h^*$. Rewards are sparse within the realized episode: $r_{h^*} = +1$ for survival and -1 for in-hospital mortality, while $r_h = 0$ for $h < h^*$.

To quantify observation patterns, we additionally define modality-specific summary statistics. For structured data, let $\boldsymbol{\delta}_h^{\mathsf{s}} \in \mathbb{R}_+^{|\mathcal{T}_h| \times D}$ denote the time since last observation for each variable at each timestamp within decision step h. For text data, let $\boldsymbol{n}_h^{\mathsf{t}} = [n_h^{\mathsf{t},j}] \in \mathbb{Z}_+^{|\mathcal{M}^{\mathsf{t}}|}$ denote the number of text observations in modality j at step h; for example, a single step may contain multiple nursing notes. We further define the documentation density

$$\kappa_h^{\mathsf{t}} = \frac{1}{K} \sum_{u=h-K+1}^{h} \sum_{j \in \mathcal{M}^{\mathsf{t}}} n_u^{\mathsf{t},j},$$

which summarizes recent documentation activity over a rolling window of length K. Together, these summaries capture behavior-driven observation timing, documentation burstiness, and *missing-not-at-random* (MNAR) patterns that may correlate with patient severity. Appendix A summarizes the notation used throughout the paper.

Learning Objectives. Given the static dataset $\mathcal{D} = \{(o_h^{(i)}, a_h^{(i)}, r_h^{(i)})\}_{i,h}$ collected under behavior policy π_β (clinician decisions), we aim to achieve three interconnected objectives:

Q1 (State Learning). Learn an encoder g_θ such that the state $s_h = g_\theta(I_h)$ captures sufficient information from multimodal observations with MNAR patterns and behavior-driven text observations. We verify state quality through reconstruction: a decoder f_ϕ should recover the step-level observations $(\boldsymbol{y}_h^{\mathsf{s}}, \boldsymbol{m}_h^{\mathsf{s}})$, ensuring that MNAR information is preserved in the learned representation.

Q2 (Policy Optimization). Under the offline constraint, learn a policy $\pi(a_h \mid s_h)$ that maximizes expected return while avoiding catastrophic errors from out-of-distribution actions.

Q3 (Outcome Prediction). Predict clinical outcomes (e.g., post-72-hour mortality) from the terminal state representation. This grounds learned states in clinically meaningful signals.

3 Method

We propose a two-stage framework as shown in Figure 2. Stage 1 (Section 3.1) learns patient health state representations from multimodal observations with structured-measurement MNAR and text MNAR; Stage 2 (Section 3.2) uses these states to optimize treatment policies via Implicit Q-Learning and to predict outcomes.

3.1 Patient State Representation Learning

To learn patient state, we adopt a *Bayesian filtering* perspective with the causal diagram shown in Figure 3. At each decision step h, we first encode the multimodal observations into a unified representation ϕ_h that explicitly captures both structured-measurement MNAR and text MNAR (Section 3.1.1). We then maintain a *latent belief state* z_h, learned via *variational inference*, whose transition to z_{h+1} is conditioned on the treatment action (Section 3.1.2). Finally, we construct the *posterior patient health state representation* s_h by combining the current observation representation ϕ_h with the latent belief state z_h (Section 3.1.3). The resulting s_h is used for downstream tasks.

3.1.1 MNAR-Aware Observation Embedding

We construct a unified representation ϕ_h through two steps: first encoding irregular structured ob-

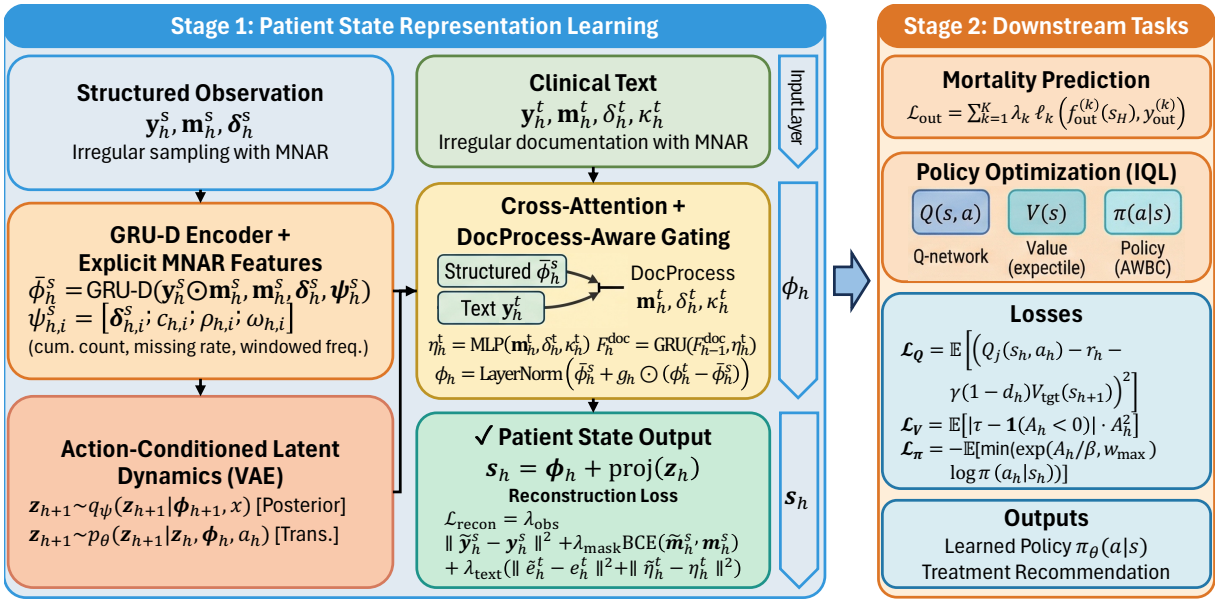

Figure 2: OPL-MT-MNAR: **Stage 1** learns state s_h with structured-measurement MNAR, documentation-process MNAR, and action-conditioned latent dynamics; **Stage 2** uses s_h for outcome prediction and policy optimization.

servations with explicit MNAR modeling to obtain $\bar{\phi}_h^s$, then incorporating clinical text together with its documentation process to produce ϕ_h.

Structured Observation Encoding. For structured observations, we use an encoder built on GRU-D (Che et al., 2018). The key idea is that if a variable has been missing for a long time, then its last observed value becomes less reliable, and the influence of stale information should gradually decay over time. To capture this effect, we introduce learned hidden-state and input decay factors at timestamp u as functions of the time-since-last-observation vector:

$$\xi_{\phi,u} = \exp\left(-\max(0, W_{\phi,\xi} \cdot \text{mean}(\delta_u) + b_{\phi,\xi})\right),$$
$$\boldsymbol{\xi}_{y,u} = \exp\left(-\max(0, W_{y,\xi}\delta_u + b_{y,\xi})\right),$$

where $\xi_{\phi,u} \in (0,1]$ controls hidden-state decay, $\boldsymbol{\xi}_{y,u} \in (0,1]^D$ controls input decay, $\xi_{y,u}^d$ denotes the d-th entry of $\boldsymbol{\xi}_{y,u}$, and all weights and biases are trainable parameters.

Let $y_u^{s,d}$ denote the value of the d-th structured variable at time u, and let $m_u^{s,d} \in \{0,1\}$ indicate whether that value is observed. When a measurement is missing, we decay its last observed value toward a default value given by the empirical mean

$$\hat{y}_u^{s,d} = m_u^{s,d} \cdot y_u^{s,d} + (1 - m_u^{s,d}) \cdot \left(\xi_{y,u}^d \cdot y_{u'}^{s,d} + (1 - \xi_{y,u}^d) \cdot \mu^d\right),$$

where $y_{u'}^{s,d}$ is the most recent observed value of the d-th variable before time u, and μ^d is the empirical mean of that variable.

Let ϕ_{u-1}^s denote the hidden state from the previous timestamp. We first apply hidden-state decay, $\hat{\phi}_{u-1}^s = \xi_{\phi,u} \odot \phi_{u-1}^s$, and then update the hidden state using GRU-D-style gates. In addition to the decayed inputs, we incorporate explicit MNAR features $\boldsymbol{\psi}_u^s$ that summarize monitoring patterns predictive of patient severity, including cumulative observation counts, missing rates, and windowed observation frequencies. The gated updates are

$$r_u = \sigma(W_{r,y}\hat{y}_u^s + W_{r,\phi}\hat{\phi}_{u-1}^s + W_{r,\psi}\boldsymbol{\psi}_u^s + b_r)$$
$$\eta_u = \sigma(W_{\eta,y}\hat{y}_u^s + W_{\eta,\phi}\hat{\phi}_{u-1}^s + W_{\eta,\psi}\boldsymbol{\psi}_u^s + b_\eta)$$
$$\tilde{\phi}_u^s = \tanh(W_y\hat{y}_u^s + W_\phi(r_u \odot \hat{\phi}_{u-1}^s) + W_\psi\boldsymbol{\psi}_u^s + b)$$
$$\phi_u^s = (1 - \eta_u) \odot \hat{\phi}_{u-1}^s + \eta_u \odot \tilde{\phi}_u^s,$$

where σ denotes the sigmoid function and all W matrices and b vectors are trainable parameters. For decision step h, we take the hidden state at the last timestamp within the step as the structured embedding $\bar{\phi}_h^s$. See Appendix D for more details.

Sparse Text Fusion. Clinical text is highly informative but is observed irregularly, and its availability is itself shaped by clinician documentation behavior. We therefore begin by introducing a *documentation-process factor* to summarize this behavior over time. At each decision step h, this factor is updated from note presence, text recency, and recent documentation density:

$$\eta_h^t = \text{MLP}\left(\boldsymbol{m}_h^t, \delta_h^t, \kappa_h^t\right)$$
$$F_h^{\text{doc}} = \text{GRU}\left(F_{h-1}^{\text{doc}}, \eta_h^t\right),$$

where η_h^{t} is a step-level summary of the text observation pattern, and F_h^{doc} accumulates these signals over time. Importantly, this documentation-process factor is constructed only from the observation process and does not directly use text content.

Next, we construct a representation of the available text at decision step h. To align textual information with the structured representation $\bar{\phi}_h^{\mathsf{s}}$, we obtain the representation ϕ_h^{t} by applying multi-head cross-attention from $\bar{\phi}_h^{\mathsf{s}}$ to embedding e_h^{t}:

$$\phi_h^{\mathsf{t}} = \text{MultiHead}\left(W_Q \bar{\phi}_h^{\mathsf{s}},\, W_K e_h^{\mathsf{t}},\, W_V e_h^{\mathsf{t}}\right).$$

Here, W_Q, W_K, and W_V are trainable projection matrices. When a text modality is unavailable, we use a learned missing embedding rather than dropping that modality entirely.

Finally, we adaptively fuse the structured representation $\bar{\phi}_h^{\mathsf{s}}$ and the text representation ϕ_h^{t} using the documentation-process factor F_h^{doc}. This allows the model to weigh text based on its semantic content and how it is documented. Specifically, we use the following gating mechanism:

$$\begin{aligned}
\hat{\phi}_h^{\mathsf{t}} &= \phi_h^{\mathsf{t}} + W_d F_h^{\text{doc}}, \\
g_h &= \sigma\big(W_g[\bar{\phi}_h^{\mathsf{s}}; \hat{\phi}_h^{\mathsf{t}}; F_h^{\text{doc}}] + b_g\big), \\
\phi_h &= \text{LayerNorm}\big(\bar{\phi}_h^{\mathsf{s}} + g_h \odot (\hat{\phi}_h^{\mathsf{t}} - \bar{\phi}_h^{\mathsf{s}})\big).
\end{aligned}$$

Here, $\hat{\phi}_h^{\mathsf{t}}$ augments the text representation with documentation-process information. The gate g_h adaptively controls how much the *unified representation* ϕ_h should move from the structured embedding toward the text-enhanced representation. In this way, the model can distinguish between settings such as no text, stale text, and a burst of newly updated text, even when the underlying text content is similar. See Appendix E for more details.

3.1.2 Latent Belief State via Variational Inference

The unified representation ϕ_h summarizes the information available at the current decision step. However, it may not fully capture the underlying health dynamics or the cumulative effects of past treatment actions. To address this limitation, we use Bayesian filtering and introduce a latent belief state z_h with action-conditioned dynamics. See Appendix K for illustrative healthcare examples.

By conditioning on action a_h, the transition from z_h to z_{h+1} captures how intervention histories shape patient trajectories and treatment responsiveness. In contrast, if z_{h+1} depends only on the previous belief state z_h and the current representation ϕ_{h+1}, but not on a_h, then the resulting model may fail to support policies that optimize long-term rewards. The key reason is that, under the causal structure in Figure 3, ϕ_{h+1} is a deterministic function of the recorded observations, which implies $\partial \phi_{h'} / \partial a_h = 0$ for all future steps $h' > h$.

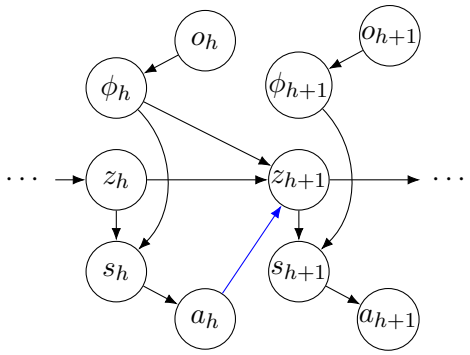

Figure 3: Causal diagram for patient state learning.

Definition 1 (Action-Independent Dynamics)**.** In the offline RL setting where $\partial \phi_{h'} / \partial a_h = 0$ (as established above), a latent dynamical system has *action-independent dynamics* if the transition function satisfies $z_{h+1} = f(z_h, \phi_h, \omega_h)$ with $\omega_h \perp\!\!\!\perp a_h$, where ω_h is exogenous noise independent of the learned policy's action.

Theorem 1 (Necessity of Action-Conditioning)**.** *Let the policy objective be* $J(\pi) = \mathbb{E}_{\tau\sim\pi}[\sum_{h=0}^{H-1} \xi^h r_h]$ *where* $r_h = R(s_h, a_h)$ *and* $s_h = g_\theta(\phi_h, z_h)$ *for a differentiable combination function* g_θ*. Under action-independent dynamics (Definition 1), the policy gradient satisfies:*

$$\frac{\partial}{\partial \pi} \mathbb{E}\left[\sum_{h'=h+1}^{H-1} \gamma^{h'} r_{h'} \;\middle|\; s_h, a_h\right] = 0 \quad \textit{for all } h.$$

That is, current actions have no gradient signal from future rewards.

With terminal-only rewards, this implies the policy gradient is zero for all non-terminal steps, making it impossible to learn that early interventions affect long-term outcomes. The proof is in Appendix J. Importantly, this theorem does not compete with MNAR-aware observation modeling: richer observation encoding improves what the model *sees* at step h, while action-conditioned latent dynamics preserve what the policy can still *learn* from future rewards.

VAE Formulation. We then parameterize z_h using a variational autoencoder (VAE):

$$z_{h+1} \sim p_\theta(z_{h+1} \mid z_h, \phi_h, a_h) = \mathcal{N}\big(\mu_\theta(z_h, \phi_h, a_h), \sigma^2_\theta(z_h, \phi_h, a_h)\big) \quad (1)$$

where μ_θ and σ_θ are parameterized by neural networks, and $z_0 \sim \mathcal{N}(0, I)$. The posterior $q_\psi(z_{h+1} \mid \phi_{h+1}, x)$ incorporates the next-step observations during training. A dynamics loss $\mathcal{L}_{\text{dyn}}$ enforces consistency between predicted and inferred states. Appendix F provides the full VAE architecture, dynamics loss, and KL regularization details.

3.1.3 Posterior Patient State Representation

The final patient health state s_h combines unified observation representation ϕ_h with the latent belief z_h via a learnable combination function:

$$s_h = g_\theta(\phi_h, z_h),$$

for a parametrized function g_θ. In our implementation, we use a residual additive form $g_\theta(\phi_h, z_h) = \phi_h + \text{proj}(z_h)$, where $\text{proj}(\cdot)$ is a linear projection. This choice preserves the representation ϕ_h while augmenting it with belief z_h that contains action-related information. Because z_h is latent and stochastic under Eq. (1), the induced state s_h is also stochastic conditional on the observation history I_h; throughout the paper, losses involving s_h are therefore understood as expectations over the corresponding latent-state draws.

State Verification via Reconstruction. To verify that the learned state captures sufficient information (Q1), we use a reconstruction objective during representation pre-training:

$$\mathcal{L}_{\text{recon}} = \lambda_{\text{obs}} \|\tilde{\boldsymbol{y}}^{\text{s}}_h - \boldsymbol{y}^{\text{s}}_h\|^2 + \lambda_{\text{mask}} \,\text{BCE}(\tilde{\boldsymbol{m}}^{\text{s}}_h, \boldsymbol{m}^{\text{s}}_h) + \lambda_{\text{text}} \Big(\|\tilde{e}^{\text{t}}_h - e^{\text{t}}_h\|^2 + \|\tilde{\eta}^{\text{t}}_h - \eta^{\text{t}}_h\|^2 \Big),$$

where $\tilde{\boldsymbol{y}}^{\text{s}}_h$, $\tilde{\boldsymbol{m}}^{\text{s}}_h$, $\tilde{e}^{\text{t}}_h$, and $\tilde{\eta}^{\text{t}}_h$ are reconstructed from s_h using MLP-based decoders. The reconstruction terms for $\tilde{\boldsymbol{y}}^{\text{s}}_h$ and $\tilde{e}^{\text{t}}_h$ encourage the state to preserve sufficient information about the structured observations and encoded text content. The term for $\tilde{\boldsymbol{m}}^{\text{s}}_h$ encourages the state to retain structured-data MNAR patterns, while the term for $\tilde{\eta}^{\text{t}}_h$ ensures that the text documentation process is also captured.

3.2 Policy Learning and Outcome Prediction

We next use the learned state representations from Stage 1 for two downstream tasks: treatment policy optimization and adverse outcome prediction.

3.2.1 Offline Policy Optimization

We adopt Implicit Q-Learning (IQL) (Kostrikov et al., 2022) for policy optimization. We first learn value functions from offline data without querying out-of-distribution actions. Next, we extract a policy via advantage-weighted behavioral cloning.

Value Function Learning. To mitigate overestimation bias, we use double Q-learning (van Hasselt, 2010) and maintain two Q-networks:

$$\mathcal{L}_{Q_j} = \mathbb{E}_{(s_h, a_h, r_h, s_{h+1}, d_h) \sim \mathcal{D}} \Big[\big(Q_j(s_h, a_h) - y^{\text{tgt}}_h \big)^2 \Big]$$

for $j \in \{1, 2\}$, where both critics are trained using the same bootstrap target

$$y^{\text{tgt}}_h = r_h + \gamma(1 - d_h) V_{\text{tgt}}(s_{h+1}),$$

$d_h \in \{0, 1\}$ indicates whether the episode terminates at step h, and V_{tgt} is a slowly updated target copy of the value function V used to stabilize bootstrapping (Appendix H). The expectation is taken over one-step transitions in the offline dataset, so the summation over decision steps is implicit in the dataset average. Value function V is trained via expectile regression (Kostrikov et al., 2022):

$$\mathcal{L}_V = \mathbb{E}_{(s_h, a_h) \sim \mathcal{D}} \Big[\big| \tau - \mathbb{I}(A_h < 0) \big| \cdot A_h^2 \Big],$$

where $A_h = \min(Q_1(s_h, a_h), Q_2(s_h, a_h)) - V(s_h)$ is the advantage and $\tau \in (0.5, 1)$ is the expectile parameter. This asymmetric loss pushes V toward higher Q-values while grounded in the data distribution.

Policy Extraction. The policy is extracted via advantage-weighted behavioral cloning:

$$\mathcal{L}_\pi = -\mathbb{E}_{(s_h, a_h) \sim \mathcal{D}} \big[\min\big(\exp(A_h / \beta), w_{\max} \big) \log \pi(a_h \mid s_h) \big],$$

where $\beta > 0$ controls deviation from clinician behavior and $w_{\max}$ prevents large weights. This formulation stays close to the behavior policy while improving on it. This is essential when a distribution shift leads to harmful recommendations.

Action Selection. We first draw the latent belief z_h from its predictive distribution, obtain the patient state $s_h = g_\theta(\phi_h, z_h)$, and then sample an action from $\pi(\cdot \mid s_h)$. Equivalently, the induced action distribution conditional on the available history is the marginal

$$\pi(a_h \mid I_h) = \int \pi(a_h \mid g_\theta(\phi_h, z_h)) \, p(z_h \mid I_h) \, dz_h,$$

which makes the uncertainty in z_h explicit while keeping the notation in the policy and value losses compact. See more details in Appendix H.

3.2.2 Outcome Prediction

Among patients who remain alive through the 72-hour observation window ($H_i = H$ and $r_H = +1$), we predict subsequent clinical outcomes $y_{\text{out}}^{(1)}, \ldots, y_{\text{out}}^{(K)}$. Our primary outcome is post-72-hour in-hospital mortality. This auxiliary target is clinically meaningful while remaining distinct from the RL reward. As a result, early-terminated episodes contribute to representation learning and policy optimization, but not to this auxiliary prediction task. We define the multitask outcome prediction loss as

$$\mathcal{L}_{\text{out}} = \sum_{k=1}^{K} \lambda_k \, \ell_k \left(f_{\text{out}}^{(k)}(s_H), y_{\text{out}}^{(k)} \right),$$

where $f_{\text{out}}^{(k)}$ denotes the prediction head for outcome k, ℓ_k is an appropriate loss function (e.g., binary cross-entropy for classification), and λ_k controls the relative weight of each outcome.

Remark 1 (Information transfer across tasks). Empirically, we observe that jointly learning policy optimization and outcome prediction improves performance on both tasks. This is because the RL reward evaluates the long-term quality of treatment decisions, whereas the auxiliary prediction task provides a more direct, less noisy supervisory signal for learning s_h. As a result, the two tasks exhibit positive transfer, yielding a patient representation s_h that is both clinically meaningful and useful for treatment optimization.

3.3 End-to-End Training Procedure

We adopt a three-stage procedure: (1) pre-train the encoder, dynamics model, and outcome predictor with reconstruction and auxiliary losses; (2) freeze the encoder and train RL components to prevent representation drift; (3) jointly fine-tune all components with reduced encoder learning rate. We monitor policy entropy to detect and recover from collapse. Appendix I reports the training schedule, loss weights, and posterior-collapse diagnostics, and Appendix R summarizes computational cost.

4 Experiments

4.1 Experimental Setup

Datasets. Following Komorowski et al. (2018), we extract sepsis cohorts from three ICU databases using a 72-hour observation window (H=18 steps at 4-hour intervals). MIMIC-III is the main benchmark: it is single-center, contains 15,415 ICU stays (13.2% mortality), and provides a high-frequency documentation regime with nursing-note coverage at 94.2% of decision steps and diagnostic-text coverage of 41.3% / 28.6% for radiology / microbiology. MIMIC-IV serves as a complementary low-frequency diagnostic-text benchmark with 32,837 ICU stays (11.8% mortality) and radiology / microbiology coverage of 83.6% and 45.5% at the step level. eICU serves as the cross-institutional, text-free generalization benchmark with 24,562 ICU stays (12.1% mortality). To prevent leakage, text is aligned so that only reports with timestamp $\leq t_h$ are available at decision step h. See full cohort construction details in Appendix B, the MIMIC-III regime breakdown in Appendix B.2, and the timestamp-alignment analysis in Appendix C.

Evaluation. Our main metric is Fitted Q-Evaluation (FQE) (Le et al., 2019), which learns a Q-function without importance weighting, critical when policies diverge from clinician behavior. Additional metrics, including Weighted Importance Sampling, are defined in Appendix L.

Baselines. We use a representative set of baselines for policy comparison: classical sepsis RL (Raghu et al., 2017b), standard offline RL (BCQ, CQL), and recent healthcare RL directions including continuous-action control (Huang et al., 2022), irregular-interval offline RL (Fatemi et al., 2022), and model-based planning (Xu et al., 2025). All methods are evaluated under their canonical action/time-step settings. On MIMIC-III, we compare structured-only modeling, low-frequency diagnostic text, nursing notes, and the full text configuration. See Appendix N for more baseline details.

4.2 Policy Learning Results

As shown in Table 1, the OPL-MT-MNAR policy reaches FQE 0.679 on MIMIC-III, compared with clinician behavior at 0.528. The gain also persists on the complementary benchmarks, reaching 0.634 on MIMIC-IV versus 0.521 for clinician behavior and 0.604 on eICU versus 0.534.

The OPL-MT-MNAR policy also improves over recent policy-learning baselines. On MIMIC-III, OPL-MT-MNAR improves over DDPG with Clinician Supervision (0.529), SBCQ (0.501), and MedDreamer (0.583); the same ranking holds on

	Method	Info	MIMIC-III Test FQE	MIMIC-IV Test FQE	eICU Test FQE
Baselines	Continuous State-Space DDQN (Raghu et al., 2017b)	Model-free	0.476	0.483	0.469
	AI Clinician (Komorowski et al., 2018)	Model-free	0.487	0.491	0.478
	BCQ (Fujimoto et al., 2019)	Model-free	0.452	0.458	0.448
	CQL (Kumar et al., 2020)	Model-free	0.411	0.418	0.408
	DDPG with Clinician Supervision (Huang et al., 2022)	Model-free	0.529	0.538	0.524
	SBCQ (Fatemi et al., 2022)	Model-free	0.501	0.508	0.494
	MedDreamer (Xu et al., 2025)	Model-based	0.583	0.591	0.579
	Clinician (Behavior)	Behavior	*0.528*	*0.521*	*0.534*
	OPL-MT-MNAR	**MNAR + Text**	**0.679**	**0.634**	**0.604**
Text Regime	OPL-MT-MNAR (Structured Only)	None	0.574	0.606	–
	OPL-MT-MNAR (Low-Freq. Text)	Rad. + Micro.	0.596	0.634	–
	OPL-MT-MNAR (Nursing Notes)	Nursing	0.624	–	–
	OPL-MT-MNAR	**All**	**0.679**	**0.634**	–

Table 1: Main policy results (test FQE). The upper block compares policies across MIMIC-III, MIMIC-IV, and eICU; OPL-MT-MNAR outperforms all baselines. The lower block shows the value of text for policy learning. The '–' entries indicate unavailable text modalities.

Configuration	What is Added	MIMIC-III FQE	Δ vs Baseline
Baseline (MDP, no MNAR)	Strong offline RL backbone	0.507	–
+ Semi-MDP	Variable-interval handling	0.518	+2.2%
+ MNAR + DocProcess (OPL-MT-MNAR)	Explicit MNAR + DocProcess	0.679	+33.9%
+ MNAR + DocProcess + Semi-MDP	MNAR + DocProcess + Semi-MDP	0.689	+35.9%

Table 2: Controlled building-block study on MIMIC-III. Explicit MNAR + DocProcess modeling provides the dominant improvement over a strong backbone; Semi-MDP handling is complementary. A broader missingness-handling encoder benchmark is reported in Table 32.

MIMIC-IV (0.538, 0.508, and 0.591 versus 0.634) and eICU (0.524, 0.494, and 0.579 versus 0.604).

Text provides substantial value for treatment policy learning. On MIMIC-III, adding low-frequency diagnostic text improves FQE from 0.574 to 0.596, incorporating nursing notes further raises it to 0.624, and the full configuration, OPL-MT-MNAR, reaches 0.679. These results show that nursing notes provide the largest single-modality gain, and the OPL-MT-MNAR configuration yields the best overall performance. Text is also valuable in MIMIC-IV, where OPL-MT-MNAR achieves FQE 0.634 in the complementary low-frequency diagnostic-text regime.

Controlled Building-Block Study. Table 2 adds the proposed components to a stronger offline RL backbone. The largest improvement comes from explicit MNAR + DocProcess modeling, while Semi-MDP handling alone yields a smaller but still complementary gain.

Clinical Interpretations. The benefits of OPL-MT-MNAR are greatest for high-acuity patients. On MIMIC-III, clinician behavior scores 0.681 in the low-SOFA group and 0.192 in the high-SOFA (> 10) group; MedDreamer reaches 0.726 and 0.296, while the OPL-MT-MNAR policy achieves 0.763 and 0.344. The widening gap in the high-severity regime is consistent with MNAR and documentation-process signals being especially informative when acuity is high. The full severity-stratified comparison is reported in Appendix P.

Robustness Checks. On MIMIC-III, FQE with bootstrap confidence intervals gives 0.679 [0.673, 0.686] for the OPL-MT-MNAR policy versus 0.528 [0.520, 0.536] for clinician behavior. Appendix M reports additional OPE estimators and chronological robustness; Appendix B.4 reports cross-dataset transfer and held-out-center generalization; Appendix O reports decision-interval and action-granularity studies; and Appendix B.6 reports the cross-disease heart-failure cohort; Appendix Q reports constrained-policy optimization, text interpretability, and time-to-deterioration analysis.

4.3 Outcome Prediction Results

We further evaluate post-72-hour in-hospital mortality prediction for patients alive at the end of the

Method	AUROC
Mean Imputation + LSTM	0.833
Forward Fill + LSTM	0.838
GRU-D (Che et al., 2018)	0.844
BRITS (Cao et al., 2018)	0.852
mTAND (Shukla and Marlin, 2021)	0.858
MedDreamer	0.867
OPL-MT-MNAR	**0.886**

Table 3: Post-72-hour mortality prediction (MIMIC-III).

72-hour window using terminal state s_H.

As shown in Table 3, the OPL-MT-MNAR encoder achieves AUROC 0.886, surpassing GRU-D (0.844), BRITS (0.852), mTAND (0.858), and MedDreamer (0.867) on MIMIC-III. This indicates that explicit modeling of measurement MNAR together with documentation-process MNAR improves representation quality beyond strong irregular-sampling encoders and world-model baselines on this later-mortality prediction task; the broader benchmark appears in Table 32.

Text provides substantial value for outcome prediction. As shown in Table 6, moving from structured-only state learning (AUROC 0.857) to nursing-note integration raises AUROC to 0.882, while the full configuration, OPL-MT-MNAR, reaches AUROC 0.886.

5 Related Work

Our work is most closely related to the growing literature on offline reinforcement learning, particularly in critical care and sepsis treatment (Komorowski et al., 2018; Raghu et al., 2017a,b, 2018; Peng et al., 2018; Huang et al., 2022; Sun and Tang, 2025). More broadly, a rich literature on off-policy learning and evaluation has developed methods for stable policy optimization and reliable value estimation under distribution shift, which are important in high-stakes medical settings (Thomas et al., 2015; Gottesman et al., 2018; Wang et al., 2018; Tang and Wiens, 2021; Kostrikov et al., 2022). Despite substantial progress (Gottesman et al., 2019; Liu et al., 2020), most existing approaches treat clinical observations as effectively complete after preprocessing and rely only on structured data. Our work contributes to this literature in three ways: (1) treating missingness as an informative signal, (2) incorporating clinical text alongside structured data, and (3) learning patient states for policy optimization rather than relying on prespecified states.

Our work is also closely related to the literature on missing data, especially MNAR settings (Little and Rubin, 2019). In structured clinical time series, methods such as GRU-D (Che et al., 2018), BRITS (Cao et al., 2018), direct missingness modeling (Lipton et al., 2016), and Raindrop (Zhang et al., 2022b) address irregular sampling and missing values. In multimodal EHR settings, methods such as MissModal (Lin and Hu, 2023), DrFuse (Yao et al., 2024), and MUSE (Wu et al., 2024) handle missing modalities. Our setting differs in that missingness is endogenously driven by unobserved factors (Xiong and Pelger, 2023; Duan et al., 2024a,b; Chen et al., 2026). Most closely related is Liang et al. (2025), which explicitly models informative missingness in multimodal EHR, but without temporal dynamics. Our work extends this direction by studying how informative missingness evolves over time across modalities and how it can be used for decision-making and outcome prediction.

Finally, our work relates to the literature on multitask learning, which seeks to exploit shared structure across related tasks (Bengio et al., 2013). In our setting, jointly learning policy optimization and outcome prediction provides a form of positive transfer. At the same time, as the number of downstream outcomes grows, negative transfer may arise, where shared representations harm accuracy for some tasks (Wu et al., 2020; Yang et al., 2025). Understanding and mitigating such interference is an important direction for future work. Recent advances in task modeling (Li et al., 2023a,b; Zhang et al., 2026b), as well as adaptive and scalable fine-tuning for individual tasks (Li et al., 2024c,d, 2025b; Zhang et al., 2026a), offer promising tools for controlling when and how information should be shared across tasks. See Appendix S for extended discussion of related work.

6 Conclusion

We introduce OPL-MT-MNAR, a framework that explicitly models temporal MNAR patterns in multimodal EHR. By combining MNAR-aware multimodal encoding, Bayesian filtering, and joint policy optimization with outcome prediction, it learns patient representations for sequential decision-making. Experiments on MIMIC-III, MIMIC-IV, and eICU show consistent gains in both off-policy optimization and outcome prediction. These results show the value of treating observation processes as informative signals for patient representation learning and off-policy clinical decision-making.

7 Limitations

Our study has several limitations that suggest directions for future work. First, like most work in clinical offline RL, our policy evaluation relies on off-policy estimation from observational data; although we report FQE with bootstrap confidence intervals in Section 4.2 and additional robustness checks in Appendix M, these results should be viewed as quantitative evidence rather than prospective validation. Second, following common practice (Komorowski et al., 2018), we discretize the continuous treatment space into 9 actions and use 4-hour decision intervals, which simplify learning and interpretation but leave finer-grained dosing and faster control horizons for future work; Appendix O quantifies this trade-off directly. Third, although we explicitly model informative missingness, unobserved factors affecting both treatment and outcomes, such as verbal communication or bedside assessments not recorded in the EHR, may still remain and could be better addressed with richer data. Finally, our experiments focus on three U.S. ICU datasets (MIMIC-III, MIMIC-IV, and eICU); Appendices B.4 and B.6 extend the analysis to transfer and cross-disease settings, Appendix Q studies constrained-policy and deterioration-prediction extensions together with text interpretability, and Appendix I reports posterior-collapse diagnostics, but broader validation across other healthcare systems remains important for deployment.

References

Denis Agniel, Isaac S. Kohane, and Griffin M. Weber. 2018. Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *BMJ*, 361:k1479.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Yitan Li, and Lei Li. 2018. Brits: bidirectional recurrent imputation for time series. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085.

Hongyu Chen, David Simchi-Levi, and Ruoxuan Xiong. 2026. Partial identification under missing data using weak shadow variables from pretrained models. *arXiv preprint arXiv:2602.16061*.

Junting Duan, Markus Pelger, and Ruoxuan Xiong. 2024a. Factor analysis for causal inference on large non-stationary panels with endogenous treatment. *Available at SSRN 4823360*.

Junting Duan, Markus Pelger, and Ruoxuan Xiong. 2024b. Target pca: Transfer learning large dimensional panel data. *Journal of Econometrics*, 244(2):105521.

Mehdi Fatemi, Mary Wu, Jeremy Petch, Walter Nelson, Stuart J Connolly, Alexander Benz, Anthony Carnicelli, and Marzyeh Ghassemi. 2022. Semi-markov offline reinforcement learning for healthcare. In *Conference on Health, Inference, and Learning*, pages 119–137. PMLR.

Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062.

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

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

Josiah P. Hanna, Peter Stone, and Scott Niekum. 2017. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*.

Milos Hauskrecht and Hamish S. F. Fraser. 2000. Planning treatment of ischemic heart disease with partially observable markov decision processes. *Artificial intelligence in medicine*, 18 3:221–44.

Yong Huang, Rui Cao, and Amir Rahmani. 2022. Reinforcement learning for sepsis treatment: A continuous action space solution. In *Proceedings of the 7th Machine Learning for Healthcare Conference*.

Nan Jiang and Lihong Li. 2016. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

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

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

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

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

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative q-learning for offline reinforcement learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Hoang Le, Cameron Voloshin, and Yisong Yue. 2019. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712.

Baohong Li, Haoxuan Li, Ruoxuan Xiong, Anpeng Wu, Fei Wu, and Kun Kuang. 2024a. Learning shadow variable representation for treatment effect estimation under collider bias. In *Forty-first International Conference on Machine Learning*.

Baohong Li, Yingrong Wang, Anpeng Wu, Ming Ma, Ruoxuan Xiong, and Kun Kuang. 2025a. Generalizing causal effects from randomized controlled trials to target populations across diverse environments. In *Forty-second International Conference on Machine Learning*.

Baohong Li, Anpeng Wu, Ruoxuan Xiong, and Kun Kuang. 2024b. Two-stage shadow inclusion estimation: An iv approach for causal inference under latent confounding and collider bias. In *Forty-first International Conference on Machine Learning*.

Dongyue Li, Haotian Ju, Aneesh Sharma, and Hongyang R Zhang. 2023a. Boosting multitask learning on graphs through higher-order task affinities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Dongyue Li, Huy L Nguyen, and Hongyang R Zhang. 2023b. Identification of negative transfers in multitask learning using surrogate models. *Transactions on Machine Learning Research*.

Dongyue Li, Aneesh Sharma, and Hongyang R Zhang. 2024c. Scalable multitask learning using gradient-based estimation of task affinity. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Dongyue Li, Ziniu Zhang, Lu Wang, and Hongyang R Zhang. 2024d. Scalable fine-tuning from multiple data sources: A first-order approximation approach. *Findings of the Association for Computational Linguistics*.

Dongyue Li, Ziniu Zhang, Lu Wang, and Hongyang R Zhang. 2025b. Efficient ensemble for fine-tuning language models on multiple datasets. *The 63rd Annual Meeting of the Association for Computational Linguistics*.

Zihan Liang, Ziwen Pan, and Ruoxuan Xiong. 2025. Causal representation learning from multimodal clinical records under non-random modality missingness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Ronghao Lin and Haifeng Hu. 2023. MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702.

Zachary C Lipton, David Kale, and Randall Wetzel. 2016. Directly modeling missing data in sequences with rnns: Improved classification of clinical time series. In *Proceedings of the 1st Machine Learning for Healthcare Conference*.

Roderick J. A. Little and Donald B. Rubin. 2019. *Statistical Analysis with Missing Data*. Wiley.

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

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. 2021. Awac: Accelerating online reinforcement learning with offline datasets.

Xuefeng Peng, Yi Ding, David Wihl, Omer Gottesman, Matthieu Komorowski, Li-wei H Lehman, Andrew Ross, Aldo Faisal, and Finale Doshi-Velez. 2018. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178.

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

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017a. Deep reinforcement learning for sepsis treatment.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. 2017b. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*.

Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. 2018. Model-based reinforcement learning for sepsis treatment.

Satya Narayan Shukla and Benjamin Marlin. 2021. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*.

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, and 1 others. 2016. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810.

Yingchuan Sun and Shengpu Tang. 2025. Exploring time-step size in reinforcement learning for sepsis treatment.

Shengpu Tang and Jenna Wiens. 2021. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*.

Philip S. Thomas and Emma Brunskill. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. 2015. High confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Hado van Hasselt. 2010. Double q-learning. In *Advances in Neural Information Processing Systems*.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. 2018. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Nicole Gray Weiskopf and Chunhua Weng. 2013. Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151.

Sen Wu, Hongyang R Zhang, and Christopher Ré. 2020. Understanding and improving information transfer in multi-task learning. *International Conference on Learning Representations*.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. 2024. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*.

Ruoxuan Xiong and Markus Pelger. 2023. Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*, 233(1):271–301.

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

Fan Yang, Hongyang R Zhang, Sen Wu, Christopher Re, and Weijie J Su. 2025. Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *Journal of Machine Learning Research*, 26(113):1–88.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. 2024. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. 2022a. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. 2022b. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*.

Zhenshuo Zhang, Minxuan Duan, Youran Ye, and Hongyang R Zhang. 2026a. Scalable multi-objective and meta reinforcement learning via gradient estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhenshuo Zhang, Minxuan Duan, and Hongyang R Zhang. 2026b. Efficient estimation of kernel surrogate models for task attribution. *arXiv preprint arXiv:2602.03783*.

Use of AI Assistants

We used an AI assistant (e.g., ChatGPT) during manuscript preparation for *language editing* and *clarity improvements* (e.g., rewriting sentences for readability and concision, suggesting alternative phrasing, and checking for grammatical consistency). The AI assistant was *not* used to generate scientific claims, derive theoretical results, design the proposed method, run experiments, perform statistical analyses, or interpret results.

All technical content—including the problem formulation, model design, implementation, experimental setup, evaluation, and conclusions—was developed, validated, and verified by the authors. We take full responsibility for the integrity, correctness, and originality of the work and for ensuring that the manuscript accurately reflects our methods and findings.

Appendix Overview

Supporting material is organized as follows: notation summary appears in Appendix A; cohort construction, transfer protocols, and the heart-failure cohort appear in Appendices B, B.4, and B.6; timestamp alignment and documentation-behavior analyses appear in Appendix C; encoder, text-fusion, latent-dynamics, outcome, IQL, and training details appear in Appendices D–I; proofs and partial-observability discussion appear in Appendices J and K; evaluation metrics and OPE robustness appear in Appendices L and M; broader baselines and ablations appear in Appendices N and O; subgroup and clinical analyses appear in Appendices P and Q; computational cost appears in Appendix R; and extended related work appears in Appendix S.

A **Notation Summary** – Symbols, indices, and key variables used throughout the paper.

B **Dataset Details** – MIMIC-III / MIMIC-IV / eICU cohort construction, cross-disease heart-failure cohort, and transfer/generalization protocols.

C **Text Timestamp Alignment** – Alignment of clinical text with decision steps, temporal buffers, gap-stratified gains, and note-behavior analyses.

D **Encoder Implementation Details** – Input construction, MNAR feature design, GRU-D backbone, and architecture specifications.

E **Text Fusion Details** – Cross-attention, documentation-process factor, DocProcess ablations, and fusion architecture.

F **Latent Dynamics Details** – Action-conditioned latent model, training losses, and KL regularization.

G **Outcome Prediction Details** – Multitask prediction heads, loss formulation, relationship to reward design, and auxiliary-vs-RL gradient-path comparison.

H **IQL Implementation Details** – Expectile regression, target network updates, advantage weighting, and hyperparameters.

I **Training Details** – Three-stage training procedure, loss weights, optimization settings, and posterior-collapse diagnostics.

J **Theoretical Analysis** – Proofs and technical discussion of controllability and multi-step credit assignment.

K **Partial Observability Discussion** – Formalization of partial observability and relation to latent state modeling.

L **Evaluation Metrics** – Definitions and estimation procedures for FQE, WIS, and auxiliary metrics.

M **Off-Policy Evaluation Details** – Bootstrap FQE, chronological split robustness, and shadow-mode evaluation notes.

N **Baseline Details** – Representative baseline families, recent sepsis RL baselines, and fairness notes.

O **Ablation Studies** – Additional controlled comparisons across temporal granularity, action granularity, and DocProcess controls.

P **Additional Subgroup Analysis** – Severity-focused breakdowns and high-acuity comparisons.

Q **Clinical Analysis Details** – Constrained policy optimization, time-to-deterioration analysis, and text interpretability.

R **Computational Analysis** – Model size, training time, and resource usage.

S **Extended Related Work** – Additional discussion of clinical RL, informative missingness, and offline RL literature.

A Notation Summary

Table 4 provides a comprehensive summary of the notation used throughout this paper. We use h to index decision steps and H to denote the horizon, following standard reinforcement learning conventions.

Dimension specifications. In our sepsis treatment experiments (Section 4): observation dimension $D = 16$ (8 vitals, 8 labs), static features $S = 3$ (age, gender, Charlson index), decision horizon $H = 18$ (4-hour intervals over 72 hours, so $T = 72$ and $\Delta = 4$), action space $|\mathcal{A}| = 9$ (3 fluid levels $\times$ 3 vasopressor levels), hidden dimension $d_h = 128$, latent dimension $d_z = 32$, and text embedding dimension $d_e = 256$.

B Dataset Details

B.1 MIMIC-IV

MIMIC-IV v2.2 (Johnson et al., 2023) is a single-center dataset from Beth Israel Deaconess Medical Center containing de-identified EHR data. We apply Sepsis-3 criteria (Singer et al., 2016): suspected infection (antibiotic administration and body fluid culture) with SOFA score ≥ 2.

Cohort Statistics. The aligned 72-hour decision cohort contains 32,837 ICU stays with 11.8% in-hospital mortality. This is the complementary low-frequency diagnostic-text benchmark used throughout the study.

Structured Variables. We extract 16 clinical variables at 1-hour resolution:

- **Vitals (8):** Heart rate, systolic BP, diastolic BP, mean BP, respiratory rate, temperature, SpO2, GCS
- **Labs (8):** Lactate, creatinine, BUN, bilirubin, platelet count, WBC, hemoglobin, glucose

Text Processing. Radiology reports (83.6% step-level coverage) and microbiology results (45.5% step-level coverage) are preprocessed by removing headers, de-identification artifacts, and normalizing abbreviations. Reports are truncated to 512 tokens and encoded using ClinicalBERT (Alsentzer et al., 2019).

Text Timestamp Alignment. To prevent information leakage, we ensure that only clinical notes available at or before each decision step are used for state encoding. Specifically, radiology and microbiology reports are aligned using their `storetime` field (the time when the report was signed and stored in the system); if `storetime` is unavailable, we fall back to `charttime`. For each decision step h at time t_h, the raw text observation set $\boldsymbol{y}_h^{\mathrm{t}}$ contains only reports with timestamps $\leq t_h$, and the step-level text embedding e_h^{t} is computed from that filtered set. The buffering analysis in Appendix C further confirms that gains persist under increasingly strict temporal exclusion windows.

B.2 MIMIC-III

MIMIC-III (Johnson et al., 2016) provides the primary high-frequency documentation benchmark in this study. We use the same 72-hour sepsis setup as in the main experiments, but the text side is dominated by nursing notes that are refreshed at most decision steps. Table 5 summarizes the regime contrast between MIMIC-III and MIMIC-IV.

Dataset	Patients	Mortality	Nursing	Rad./Micro.
MIMIC-III	15,415	13.2%	94.2%	41.3% / 28.6%
MIMIC-IV	32,837	11.8%	N/A	83.6% / 45.5%

Table 5: Complementary text regimes in the two MIMIC cohorts. MIMIC-III supplies high-frequency nursing documentation, while MIMIC-IV emphasizes lower-frequency diagnostic text.

This regime difference is why MIMIC-III is useful beyond being an additional benchmark: it exposes a setting where documentation behavior is visible at nearly every decision step, making text-process MNAR especially easy to analyze.

Policy Performance by Text Type. Table 6 summarizes how policy value changes as we compare structured-only inputs, low-frequency diagnostic text, nursing notes, and the full text configuration.

Text Modality	Coverage	FQE	AUROC	Δ vs Struct.
Structured Only	0%	0.574	0.857	–
Low-Freq. Text	52.1%	0.596	0.869	+3.8%
Nursing Notes	94.2%	0.624	0.882	+8.7%
Full Text	96.4%	0.679	0.886	+18.3%

Table 6: MIMIC-III text-regime analysis. Nursing notes provide a strong single-modality gain, while the full configuration, OPL-MT-MNAR, achieves the best overall performance.

B.3 eICU

The eICU Collaborative Research Database (Pollard et al., 2018) is used as the cross-institutional

Symbol	Description
Time Indices	
$t \in \{0, \dots, T\}$	Observation time index (fine-grained grid)
$h \in \{0, \dots, H-1\}$	Decision step index
H	Decision horizon (number of decision steps)
T	Observation horizon; $T = \Delta \cdot H$ for interval Δ
Structured Observations (per decision step h)	
$\mathcal{T}_h$	Observation times contained in decision step h
$\boldsymbol{y}_h^{\mathsf{s}} \in \mathbb{R}^{\lvert\mathcal{T}_h\rvert \times D}$	Structured measurement matrix in step h
$\boldsymbol{m}_h^{\mathsf{s}} \in \{0,1\}^{\lvert\mathcal{T}_h\rvert \times D}$	Structured observation-mask matrix in step h
$\boldsymbol{\delta}_h^{\mathsf{s}} \in \mathbb{R}_+^{\lvert\mathcal{T}_h\rvert \times D}$	Structured time-gap matrix in step h
$y_{h,i}^{\mathsf{s}}, m_{h,i}^{\mathsf{s}}, \delta_{h,i}^{\mathsf{s}}$	Row-i structured value, mask, and time-gap vectors inside step h
$\boldsymbol{\psi}_t^{\mathsf{s}} \in \mathbb{R}^{4D}$	Row-level explicit MNAR features used inside the structured encoder (Appendix D)
$x \in \mathbb{R}^S$	Static patient features
$y_{\text{out}}^{(k)} \in \{0,1\}$	Clinical outcome labels (e.g., post-72-hour mortality)
Text Observations (per decision step h)	
$\boldsymbol{y}_h^{\mathsf{t}}$	Collection of raw text observations available at step h across text modalities
$e_h^{\mathsf{t}} \in \mathbb{R}^{d_e}$	Step-level text embedding encoded from $\boldsymbol{y}_h^{\mathsf{t}}$
$e_h^{\mathsf{t,r}}, e_h^{\mathsf{t,m}} \in \mathbb{R}^{d_e}$	Radiology / microbiology modality embeddings
$\boldsymbol{n}_h^{\mathsf{t}} \in \mathbb{Z}_+^{\lvert\mathcal{M}_{\text{text}}\rvert}$	Number of text observations available for each text modality in step h
$m_h^{\mathsf{t}} \in \{0,1\}^{\lvert\mathcal{M}_{\text{text}}\rvert}$	Derived text-availability indicators: $m_h^{\mathsf{t}} = \mathbf{1}[\boldsymbol{n}_h^{\mathsf{t}} > 0]$
$\delta_h^{\mathsf{t}} \in \mathbb{R}_+$	Derived text recency summary from the availability history
$\kappa_h^{\mathsf{t}} \in \mathbb{R}_+$	Derived average number of text updates per step over the previous K decision steps
Information Set	
o_h	Decision-step record containing primitive observations $(\boldsymbol{y}_h^{\mathsf{s}}, \boldsymbol{m}_h^{\mathsf{s}}, \boldsymbol{y}_h^{\mathsf{t}}, \boldsymbol{m}_h^{\mathsf{t}})$ and explicit derived recency / density features $(\boldsymbol{\delta}_h^{\mathsf{s}}, \delta_h^{\mathsf{t}}, \kappa_h^{\mathsf{t}})$
I_h	Information set up to step h: $I_h = \{x, o_1, \dots, o_h\}$
Latent Representations	
$\tilde{\phi}_h^{\mathsf{s}} \in \mathbb{R}^{d_h}$	Decision-aligned structured embedding
$\phi_h^{\mathsf{t}} \in \mathbb{R}^{d_h}$	Text-content representation before gated fusion
η_h^{t}	Documentation-process embedding from presence / recency / density
F_h^{doc}	Temporal documentation-process factor
$\phi_h \in \mathbb{R}^{d_h}$	Text-fused observation embedding
$z_h \in \mathbb{R}^{d_z}$	Latent belief state (prior)
$g_\theta(\cdot, \cdot)$	Learnable state combination function
$s_h \in \mathbb{R}^{d_s}$	Full decision state: $s_h = g_\theta(\phi_h, z_h)$
Actions and Rewards	
$a_h \in \mathcal{A}$	Discrete treatment action
$r_h \in \mathbb{R}$	Reward (terminal-only: $r_H \in \{-1, +1\}$)
$\gamma \in [0, 1)$	Discount factor
Policy and Value Functions	
$\pi_\theta(a \mid s)$	Learned policy
π_β	Behavior policy (clinician decisions)
$Q(s, a)$	State-action value function
$V(s)$	State value function
$A(s, a)$	Advantage: $A = Q - V$
Key Hyperparameters	
τ	Expectile for IQL value learning
β	Temperature for advantage-weighted policy

Table 4: Summary of notation used throughout the paper.

benchmark. Sepsis patients are identified using APACHE IV admission diagnosis codes for sepsis with SOFA $\geq$ 2. This dataset contains only structured observations without clinical text, enabling evaluation of our MNAR-aware encoder under cross-institutional shift and held-out-center generalization. Table 7 collects the in-distribution and transfer settings across MIMIC-III, MIMIC-IV, and eICU.

B.4 Cross-Dataset Validation Protocol

We evaluate cross-institutional generalization through multiple training protocols.

Protocol Details.

- **In-distribution (rows 1–3):** Models trained and tested on the same dataset. These correspond to the primary results in Section 4.2. The eICU model is trained from scratch without text fusion.
- **Zero-shot transfer (rows 4–9):** Models trained on one dataset and directly applied to the others without any adaptation. This tests whether learned representations and policies generalize across institutions.
- **Fine-tuned transfer (rows 10–15):** Models pre-trained on one dataset, then fine-tuned on the target dataset. This tests whether pre-training provides useful initialization.

Key Findings.

- **Training from scratch is best:** In-distribution models outperform transfer variants, suggesting that institution-specific patterns are important for optimal performance.
- **Zero-shot transfer improves over clinicians:** Even without any target-domain training, transferred models improve over clinician baselines on both directions of transfer, showing that the learned state captures cross-hospital treatment signal rather than only site-specific heuristics.
- **Fine-tuning narrows the gap:** Fine-tuned models partially close the distance between zero-shot and in-distribution performance, indicating that the transferred encoder is a useful initialization rather than a brittle source-only model.
- **Observation-process features transfer:** The transfer gains confirm that MNAR-aware representations generalize across institutions with different monitoring protocols.

B.5 Within-eICU Center-Held-Out Generalization

Table 8 reports held-out-center performance inside eICU, isolating geographic shift without mixing it with cross-dataset transfer.

B.6 Heart Failure Cross-Disease Validation

We evaluate a MIMIC-III heart-failure cohort with a distinct action space (Diuretic $\times$ Vasoactive) and a similarly high-frequency documentation regime, with nursing-note coverage of 93.8% at the decision-step level. Table 9 shows that the text benefit persists in this cross-disease setting.

B.7 Action Space and Temporal Granularity

Following Komorowski et al. (2018), we discretize treatments into a 3×3 action space:

- **IV fluids:** None (0), Low ($<$500 mL/4h), High ($\geq$500 mL/4h)
- **Vasopressors:** None (0), Low ($<$0.1 mcg/kg/min norepinephrine equivalent), High ($\geq$0.1 mcg/kg/min)

This yields 9 discrete actions representing clinically meaningful treatment intensities for sepsis. For the heart-failure cohort, we analogously use a Diuretic $\times$ Vasoactive grid.

Action Granularity Comparison. Table 10 compares the retained 3×3 action space against finer and continuous alternatives under the same MIMIC-III setup.

Method	Setting	MIMIC-III FQE	Action Space
DDPG with Clinician Supervision	Continuous-control baseline	0.529	Continuous
AI Clinician	Canonical tabular baseline	0.487	25 discrete
OPL-MT-MNAR	Continuous-head ablation	0.611	Continuous
OPL-MT-MNAR	5×5 discretization	0.668	25 discrete
OPL-MT-MNAR	**Main setting**	**0.679**	**9 discrete**

Table 10: Action-granularity comparison on MIMIC-III. Finer or continuous action spaces remain viable, but the 3×3 discretization gives the best value under current data coverage and OPE stability.

Decision-Interval Sweep. Table 11 complements the action-space check by sweeping decision intervals, showing that 4 hours yields the best overall trade-off in our experiments.

Training Protocol	Train Data	Test Data	Test FQE	Δ vs Clin.	AUROC
In-Distribution Evaluation					
MIMIC-III $\rightarrow$ MIMIC-III	MIMIC-III	MIMIC-III	0.679	+28.6%	0.886
MIMIC-IV $\rightarrow$ MIMIC-IV	MIMIC-IV	MIMIC-IV	0.634	+21.7%	0.879
eICU $\rightarrow$ eICU	eICU	eICU	0.604	+13.1%	0.862
Cross-Dataset Transfer (Zero-Shot)					
MIMIC-III $\rightarrow$ MIMIC-IV	MIMIC-III	MIMIC-IV	0.573	+10.0%	0.847
MIMIC-III $\rightarrow$ eICU	MIMIC-III	eICU	0.562	+5.2%	0.839
MIMIC-IV $\rightarrow$ MIMIC-III	MIMIC-IV	MIMIC-III	0.559	+5.9%	0.846
MIMIC-IV $\rightarrow$ eICU	MIMIC-IV	eICU	0.568	+6.4%	0.844
eICU $\rightarrow$ MIMIC-III	eICU	MIMIC-III	0.551	+4.4%	0.841
eICU $\rightarrow$ MIMIC-IV	eICU	MIMIC-IV	0.556	+6.7%	0.851
Cross-Dataset Transfer (Fine-tuned)					
MIMIC-III $\rightarrow$ MIMIC-IV (FT)	MIMIC-III + MIMIC-IV	MIMIC-IV	0.598	+14.8%	0.857
MIMIC-III $\rightarrow$ eICU (FT)	MIMIC-III + eICU	eICU	0.585	+9.6%	0.847
MIMIC-IV $\rightarrow$ MIMIC-III (FT)	MIMIC-IV + MIMIC-III	MIMIC-III	0.621	+17.6%	0.864
MIMIC-IV $\rightarrow$ eICU (FT)	MIMIC-IV + eICU	eICU	0.594	+11.2%	0.857
eICU $\rightarrow$ MIMIC-III (FT)	eICU + MIMIC-III	MIMIC-III	0.607	+15.0%	0.853
eICU $\rightarrow$ MIMIC-IV (FT)	eICU + MIMIC-IV	MIMIC-IV	0.619	+18.8%	0.872

Table 7: Cross-dataset validation protocols. In-distribution training remains strongest, but zero-shot transfer still improves over clinician behavior and fine-tuning recovers part of the distribution-shift gap.

Split	# Train Centers	# Test Centers	Test FQE	Δ vs Clin.	AUROC
Subsampling Split 1	146	62	0.597	+10.8%	0.858
Subsampling Split 2	146	62	0.601	+11.4%	0.860
Subsampling Split 3	146	62	0.594	+9.6%	0.857
Subsampling Split 4	146	62	0.608	+13.0%	0.863
Subsampling Split 5	146	62	0.599	+10.9%	0.859
Mean $\pm$ SD	–	–	**0.600 $\pm$ 0.005**	**+11.1% $\pm$ 1.2%**	**0.859 $\pm$ 0.002**

Table 8: Within-eICU held-out-center generalization across five independent 70%/30% center splits. The table mean of 0.600 ± 0.005 summarizes these five subsampling splits, while a separate leave-one-center-out analysis on the largest 10 centers gives a similar mean FQE of 0.597 ± 0.005. Together these two protocols indicate that observation-process features transfer more reliably than site-specific policy heads.

Dataset	Disease	Action Space	Note Cov.	Structured	Nursing Notes	Full Text	AUROC
MIMIC-III	Sepsis	Fluid $\times$ Vasopressor	94.2%	0.574	0.624	0.679	0.886
MIMIC-III	Heart Failure	Diuretic $\times$ Vasoactive	93.8%	0.557	0.597	0.603	0.864

Table 9: Cross-disease validation in the high-frequency MIMIC-III regime. Text integration improves policy value in both sepsis and heart failure, and the full heart-failure model remains predictive of post-72-hour mortality risk.

Δt	Cohort	Avg. Ep.	MIMIC-III	eICU	AUROC	Train (h)	ESS
1h	14,831	48.7	0.661	0.587	0.872	6.3	3.9
2h	15,178	24.4	0.673	0.600	0.879	3.4	4.7
4h	15,415	12.2	0.679	0.604	0.886	2.1	5.8
8h	14,267	6.3	0.658	0.589	0.876	1.4	7.4

Table 11: Decision-interval sweep on MIMIC-III. The standard 4-hour setting offers the best overall trade-off between policy value, auxiliary prediction quality, and OPE reliability.

C Text Timestamp Alignment and Sensitivity Analysis

This appendix provides detailed analysis of text timestamp alignment, text recency controls, and note-behavior mechanisms under both low-frequency (MIMIC-IV) and high-frequency (MIMIC-III) text regimes.

C.1 Timestamp Alignment Protocol

For MIMIC-III nursing notes, we align text conservatively so that only notes available at or before each decision step are used in state encoding. At each decision step h occurring at time t_h, only nursing notes with timestamp $\leq t_h$ are included in the text branch.

Coverage Impact. This temporal filtering yields the 94.2% step-level nursing-note coverage reported in Section 4.1. The buffering controls below show that gains degrade smoothly as progressively fresher notes are excluded.

C.2 Sensitivity Analysis with Temporal Buffers

To assess whether gains depend on potentially delayed reports, we evaluate performance with increasingly conservative temporal buffers that exclude reports near decision boundaries. Table 12 reports this buffering analysis directly.

Note Alignment (MIMIC-III)	Coverage	FQE	Δ
All available	94.2%	0.679	–
Conservative (−2h)	81.7%	0.672	−1.0%
Strict (−4h)	68.9%	0.661	−2.7%
Very strict (−8h)	50.4%	0.645	−5.0%
No text	0%	0.574	−15.5%

Table 12: Sensitivity analysis for MIMIC-III nursing-note alignment. Gains persist under increasingly conservative temporal buffers, supporting the claim that improvements are not driven by leakage from future notes.

Key Findings.

- **Robust to conservative alignment:** Even under an 8-hour exclusion window, the model remains above the structured-only baseline.
- **Graceful degradation:** Performance degrades smoothly as text availability decreases, consistent with a learned recency-sensitive fusion mechanism rather than a brittle dependence on the newest note.

C.3 Gap-Stratified Text Gains in the High-Frequency Regime

Table 13 quantifies where nursing notes help most as structured observation gaps widen, and Table 14 shows the corresponding shift in modality attention.

Time Gap	Patient-Steps	Struct.	+ Nursing	Δ
$\delta t \leq 1$h	31.8%	0.612	0.638	+4.2%
1h $< \delta t \leq 2$h	27.4%	0.584	0.629	+7.7%
2h $< \delta t \leq 4$h	24.1%	0.551	0.617	+12.0%
$\delta t >$ 4h	16.7%	0.498	0.604	+21.3%

Table 13: Gap-stratified nursing-note gains on MIMIC-III. Text becomes most valuable when structured observations are stale or sparse.

Time Gap	Structured	Nursing	Rad./Micro.
$\delta t \leq 1$h	0.64	0.21	0.15
$\delta t >$ 4h	0.22	0.58	0.20

Table 14: Attention shifts toward text as structured measurements become stale in the MIMIC-III nursing-note regime.

C.4 Documentation Behavior by Shift

Table 15 summarizes how note frequency and note content differ between day and night shifts in the MIMIC-III nursing-note regime.

Metric	Day Shift	Night Shift	Night/Day
% of total notes	68.4%	31.6%	0.46
Avg. note length (tokens)	142	187	1.32
Routine assessment	45.2%	18.4%	0.41
Acute status change	8.3%	24.6%	2.96
Hemodynamic instability	6.7%	19.2%	2.87
Respiratory event	5.4%	14.8%	2.74

Table 15: Shift-level nursing-note behavior on MIMIC-III. Fewer notes are written overnight, but they are longer and much more likely to describe acute deterioration.

Although night shift contributes fewer notes overall, night notes are longer and 2.7–3.0× more likely to describe acute deterioration events, supporting the view that note presence, timing, and length form a behavior-driven observation process.

C.5 Content Evolution Categories

Table 16 groups nursing-note updates by content evolution category and shows that the gate is highest when the note signals active clinical change.

The majority of reports (68–72%) are available well before the decision step (>4h), indicating that

Category	% Steps	FQE	Treatment Change	Gate	Example Keywords
First Note	27.2%	0.636	48.4%	0.50	–
Worsening	13.8%	0.668	64.7%	0.79	worsening, progressive, new infiltrate
New Critical Finding	9.2%	0.671	67.3%	0.78	new consolidation, positive culture, abscess
Improving	11.7%	0.643	38.9%	0.53	improved, resolving, clearing
Stable / Unchanged	38.1%	0.628	26.3%	0.41	stable, unchanged, no interval change

Table 16: Content-evolution analysis for nursing notes. The gate is highest when notes describe active clinical change, consistent with text acting as an observation channel for latent dynamics rather than a static side input.

most text information reflects prior clinical assessments rather than contemporaneous findings. This temporal separation provides additional confidence that text fusion does not introduce lookahead bias.

D Encoder Implementation Details

This appendix provides the complete specification of the GRU-D encoder with explicit MNAR feature fusion.

D.1 Input Construction

At each fine-grained observation time t, let $\boldsymbol{y}_t^{\mathsf{s}}, \boldsymbol{m}_t^{\mathsf{s}}, \boldsymbol{\delta}_t^{\mathsf{s}} \in \mathbb{R}^D$ denote the row-level structured value, mask, and time-gap vectors, corresponding to one row of $(\boldsymbol{y}_h^{\mathsf{s}}, \boldsymbol{m}_h^{\mathsf{s}}, \boldsymbol{\delta}_h^{\mathsf{s}})$ in the main text. Let $\boldsymbol{y}_{t'}^{\mathsf{s}}$ denote the most recent previously observed value for each variable before time t, and let $\boldsymbol{\mu}^{\mathsf{s}}$ denote the empirical-mean vector. The decayed structured input is

$$\hat{\boldsymbol{y}}_t^{\mathsf{s}} = \boldsymbol{m}_t^{\mathsf{s}} \odot \boldsymbol{y}_t^{\mathsf{s}} + (1 - \boldsymbol{m}_t^{\mathsf{s}}) \odot \big(\boldsymbol{\xi}_{y,t} \odot \boldsymbol{y}_{t'}^{\mathsf{s}} + (1 - \boldsymbol{\xi}_{y,t}) \odot \boldsymbol{\mu}^{\mathsf{s}}\big),$$

where $\boldsymbol{\xi}_{y,t}$ is the learned input-decay vector.

D.2 Explicit MNAR Features

Beyond implicit missingness modeling through $\boldsymbol{m}_t^{\mathsf{s}}$ and $\boldsymbol{\delta}_t^{\mathsf{s}}$, we compute explicit MNAR features $\boldsymbol{\psi}_t^{\mathsf{s}} \in \mathbb{R}^{4D}$ that capture observation patterns:

$$\boldsymbol{\psi}_t^{\mathsf{s}} = \big[\delta_t;\ c_t;\ \rho_t;\ \omega_t\big],$$

where:

- $\boldsymbol{\delta}_t^{\mathsf{s}} \in \mathbb{R}^D$: time since last observation for each variable
- $\boldsymbol{c}_t^{\mathsf{s}} = \sum_{\tau \le t} \boldsymbol{m}_\tau^{\mathsf{s}} \in \mathbb{R}^D$: cumulative observation count
- $\boldsymbol{\rho}_t^{\mathsf{s}} = 1 - \boldsymbol{c}_t^{\mathsf{s}}/t \in \mathbb{R}^D$: cumulative missing rate
- $\boldsymbol{\omega}_t^{\mathsf{s}} \in \mathbb{R}^D$: windowed observation frequency over the past W hours (we use $W = 6$)

These features explicitly encode the monitoring intensity that correlates with patient severity. Cumulative counts reveal overall monitoring intensity, missing rates track deterioration trends, and windowed frequency captures recent clinical concern.

D.3 GRU-D with Temporal Decay

We employ GRU-D (Che et al., 2018) with learned temporal decay. The decay factors are:

$$\xi_{\phi,t} = \exp\big(-\max(0, W_{\phi,\xi} \cdot \bar{\delta}_t + b_{\phi,\xi})\big),$$
$$\boldsymbol{\xi}_{y,t} = \exp\big(-\max(0, W_{y,\xi} \boldsymbol{\delta}_t^{\mathsf{s}} + b_{y,\xi})\big),$$

where $\bar{\delta}_t = \mathrm{mean}(\boldsymbol{\delta}_t^{\mathsf{s}})$, $\xi_{\phi,t} \in (0, 1]$ controls hidden-state decay, and $\boldsymbol{\xi}_{y,t} \in (0, 1]^D$ controls input decay toward the empirical mean $\boldsymbol{\mu}^{\mathsf{s}}$.

The hidden state update then follows the same GRU-D-style gating equations as in the main text, with the explicit MNAR features entering through a dedicated branch:

$$\begin{aligned}
\hat{\phi}_{t-1}^{\mathsf{s}} &= \xi_{\phi,t} \odot \phi_{t-1}^{\mathsf{s}}, \\
(\boldsymbol{\psi}_t^{\mathsf{s}})^{\text{emb}} &= \mathrm{MLP}_\psi(\boldsymbol{\psi}_t^{\mathsf{s}}), \\
r_t &= \sigma\Big(W_{r,y} \hat{\boldsymbol{y}}_t^{\mathsf{s}} + W_{r,\phi} \hat{\phi}_{t-1}^{\mathsf{s}} + W_{r,\psi} (\boldsymbol{\psi}_t^{\mathsf{s}})^{\text{emb}} + b_r\Big), \\
\eta_t &= \sigma\Big(W_{\eta,y} \hat{\boldsymbol{y}}_t^{\mathsf{s}} + W_{\eta,\phi} \hat{\phi}_{t-1}^{\mathsf{s}} + W_{\eta,\psi} (\boldsymbol{\psi}_t^{\mathsf{s}})^{\text{emb}} + b_\eta\Big), \\
\tilde{\phi}_t^{\mathsf{s}} &= \tanh\Big(W_y \hat{\boldsymbol{y}}_t^{\mathsf{s}} + W_\phi \big(r_t \odot \hat{\phi}_{t-1}^{\mathsf{s}}\big) + W_\psi (\boldsymbol{\psi}_t^{\mathsf{s}})^{\text{emb}} + b\Big), \\
\phi_t^{\mathsf{s}} &= (1 - \eta_t) \odot \hat{\phi}_{t-1}^{\mathsf{s}} + \eta_t \odot \tilde{\phi}_t^{\mathsf{s}}.
\end{aligned}$$

The decision-aligned embedding at step h is obtained by selecting the hidden state at the last fine-grained timestamp within that decision step:

$$\bar{\phi}_h^{\mathsf{s}} = \phi_{t^\star(h)}^{\mathsf{s}},$$

where $t^\star(h) = \max \mathcal{T}_h$.

D.4 Architecture Specifications

Table 17 lists the encoder hyperparameters used for the GRU-D backbone and explicit MNAR-feature branch.

Component	Specification
GRU-D hidden dimension	$d_h = 128$
$\boldsymbol{\delta}_t^{\mathsf{s}}$ embedding dimension	$d_\delta = 16$
MNAR feature embedding	MLP: $4D \to 32 \to 32$
Dropout rate	0.1

Table 17: Encoder architecture hyperparameters.

E Text Fusion Details

Clinical text is encoded using ClinicalBERT (Alsentzer et al., 2019). The same fusion module is used for MIMIC-IV diagnostic text and MIMIC-III nursing notes; what changes across datasets is the documentation regime, not the fusion logic.

E.1 Cross-Attention Mechanism

The structured embedding $\bar{\phi}_h^{\mathsf{s}}$ queries the encoded text embeddings via multi-head cross-attention:

$$\phi_h^{\mathsf{t}} = \mathrm{MultiHead}\left(W_Q \bar{\phi}_h^{\mathsf{s}},\, W_K e_h^{\mathsf{t}},\, W_V e_h^{\mathsf{t}}\right).$$

where e_h^{t} concatenates the modality-specific text embeddings encoded from the raw text observations $\boldsymbol{y}_h^{\mathsf{t}}$ at step h (radiology, microbiology, and nursing notes when present). If more than one nursing note or report is available within the same decision step, all such embeddings are included in e_h^{t} before attention. The notation e_h^{t} in the main text refers to this encoded step-level text input. When a text modality is unavailable, we use a learned missing embedding rather than dropping the modality entirely.

E.2 Documentation-Process Factor

Text fusion depends on both *content* and the *documentation process*. We define

$$\eta_h^{\mathsf{t}} = \mathrm{MLP}\left(m_h^{\mathsf{t}}, \delta_h^{\mathsf{t}}, \kappa_h^{\mathsf{t}}\right),$$
$$F_h^{\mathrm{doc}} = \mathrm{GRU}\left(F_{h-1}^{\mathrm{doc}}, \eta_h^{\mathsf{t}}\right),$$

where $m_h^{\mathsf{t}} = \mathbf{1}[\boldsymbol{n}_h^{\mathsf{t}} > 0]$ captures note presence derived from the step-level text counts, δ_h^{t} is a derived summary of time since the last note, and κ_h^{t} is the recent note-update density. Equation above defines only the documentation-process branch: it does not directly encode text content. The semantic content enters separately through the cross-attention path $e_h^{\mathsf{t}} \to \phi_h^{\mathsf{t}}$, and the final text contribution is formed only after ϕ_h^{t} is combined with F_h^{doc} in the gate below. These inputs summarize behavior-driven freshness and burstiness rather than explicit wall-clock covariates such as hour-of-day or shift. The gate then becomes

$$\hat{\phi}_h^{\mathsf{t}} = \phi_h^{\mathsf{t}} + W_d F_h^{\mathrm{doc}},$$
$$g_h = \sigma\left(W_g[\bar{\phi}_h^{\mathsf{s}}; \hat{\phi}_h^{\mathsf{t}}; F_h^{\mathrm{doc}}] + b_g\right),$$
$$\phi_h = \mathrm{LayerNorm}\left(\bar{\phi}_h^{\mathsf{s}} + g_h \odot (\hat{\phi}_h^{\mathsf{t}} - \bar{\phi}_h^{\mathsf{s}})\right).$$

This is the main distinction from the prior content-only gate and also from the structured branch: structured MNAR features such as $(\boldsymbol{\psi}_t^{\mathsf{s}})^{\mathrm{emb}}$ are integrated locally into the same GRU-D encoder as the measured values, whereas the text side separates semantic content from the documentation process and uses F_h^{doc} as a distinct recurrent factor that accumulates documentation behavior across decision steps before fusion.

E.3 DocProcess Component Ablation

Table 18 isolates the contribution of the documentation-process embedding, GRU factor, and individual text-process features.

E.4 Gate Mechanism Checks

Table 19 checks whether the learned gate uses text in the intended way: it should trust text more when documentation is fresh or dense, and down-weight it when documentation is stale.

E.5 DocProcess Across Text Regimes

Table 20 compares how much of the total text gain is attributable to the documentation-process factor under the low-frequency MIMIC-IV regime versus the high-frequency MIMIC-III regime.

E.6 Text Fusion Architecture

Table 21 summarizes the final text branch after adding the documentation-process factor, making the appendix architecture description consistent with the notation and equations in Section 3.1.1.

Component	Specification
Text encoder	ClinicalBERT (frozen)
Text embedding dimension	$d_e = 256$ (projected from 768)
Cross-attention heads	4
Cross-attention dimension	128
DocProcess MLP	$\{m_h^{\mathsf{t}}, \delta_h^{\mathsf{t}}, \kappa_h^{\mathsf{t}}\} \to d_h$
DocProcess factor	GRU with hidden size d_h
Gate MLP	Linear: $3d_h \to d_h$

Table 21: Text fusion architecture hyperparameters after adding the documentation-process factor.

Configuration	MIMIC-III FQE	MIMIC-III AUROC	ΔFQE
Full model	0.679	0.886	–
w/o DocProcess GRU factor	0.671	0.881	-1.2%
w/o DocProcess embedding entirely (content-only design)	0.665	0.878	-2.1%
w/o note presence indicator	0.672	0.882	-1.0%
w/o time since last note	0.670	0.881	-1.3%
w/o all MNAR features	0.548	0.839	-19.3%

Table 18: DocProcess ablation on MIMIC-III. The documentation-process embedding improves on the prior content-only design, and removing all MNAR signals is substantially more damaging than removing any single text-process feature.

Documentation Regime				Correlation With Clinical Signal		
Regime	**Gate (w/ F_{doc})**	**Gate (w/o F_{doc})**	Δ	**Statistic**	**Value**	p
High-freq. density > 0.6	0.74	0.64	+15.6%	Corr(F_{doc}, gate)	$r = 0.67$	< 0.001
Low-freq. density < 0.2	0.35	0.43	-18.6%	Corr(F_{doc}, mortality)	$r = 0.59$	< 0.001
Fresh note ($\delta_h^{\text{t}} \leq 2$h)	0.81	0.70	+15.7%	Corr(F_{doc}, FQE gain)	$r = 0.63$	< 0.001
Stale note ($\delta_h^{\text{t}} > 8$h)	0.24	0.34	-29.4%	–	–	–
Night shift (7pm–7am)	0.71	0.64	+10.9%	–	–	–

Table 19: Mechanistic checks for the documentation-process factor on MIMIC-III. The gate trusts text more when notes are fresh or dense, less when they are stale, and the learned factor correlates with both acuity and downstream policy gain.

Dataset	Full vs w/o DocProcess	Full vs Structured Only	DocProcess / Total
MIMIC-III (Nursing + Diagnostic Text)	+2.1%	+18.3%	11.5%
MIMIC-IV (Rad. + Micro.)	+1.1%	+4.7%	23.4%

Table 20: The documentation-process factor contributes a larger fraction of the text gain in the lower-frequency MIMIC-IV regime, where distinguishing fresh from stale notes is most critical.

F Latent Dynamics Details

This appendix provides details on the action-conditioned latent dynamics model.

F.1 VAE Architecture

The prior network predicts the next latent state given current state, current observation embedding, and action:

$$\mu_\theta, \log \sigma_\theta^2 = \text{MLP}_\theta\big([z_h; \phi_h; \text{Emb}(a_h)]\big),$$

where $\text{Emb}(\cdot)$ is an action embedding layer mapping discrete actions to continuous vectors.

The posterior network incorporates the next-step observations:

$$\mu_\psi, \log \sigma_\psi^2 = \text{MLP}_\psi\big([\phi_{h+1}; x]\big).$$

Both networks output mean and log-variance for a diagonal Gaussian distribution.

F.2 Dynamics Loss

The dynamics loss $\mathcal{L}_{\text{dyn}}$ enforces consistency between the prior and posterior:

$$\mathcal{L}_{\text{dyn}} = \mathbb{E}_{q_\psi}\left[\|z_{h+1} - \hat{z}_{h+1}\|^2\right],$$

where $\hat{z}_{h+1} = \mu_\theta(z_h, \phi_h, a_h)$ is the predicted mean from the prior.

F.3 KL Regularization

The KL regularization loss encourages the posterior to stay close to the prior:

$$\mathcal{L}_{\text{KL}} = \text{KL}\Big(q_\psi(z_{h+1} \mid \phi_{h+1}, x) \,\|\, p_\theta(z_{h+1} \mid z_h, \phi_h, a_h)\Big).$$

For diagonal Gaussians, this has a closed-form solution:

$$\mathcal{L}_{\text{KL}} = \frac{1}{2}\sum_{j=1}^{d_z}\left(\log\frac{\sigma_{\theta,j}^2}{\sigma_{\psi,j}^2} + \frac{\sigma_{\psi,j}^2}{\sigma_{\theta,j}^2} + \frac{(\mu_{\psi,j} - \mu_{\theta,j})^2}{\sigma_{\theta,j}^2} - 1\right).$$

F.4 Architecture Specifications

Table 22 summarizes the latent-dynamics architecture corresponding to the prior, posterior, and state-combination modules above.

Component	Specification
Latent dimension	$d_z = 32$
Action embedding dimension	16
Prior MLP	$[d_z + d_h + 16] \to 128 \to 64 \to 2d_z$
Posterior MLP	$[d_h + S] \to 128 \to 64 \to 2d_z$
Combination g_θ	Residual: ϕ_h + Linear(z_h), where Linear: $d_z \to d_h$

Table 22: Latent dynamics architecture hyperparameters.

G Outcome Prediction Details

G.1 Relationship Between Outcomes and Rewards

For the primary auxiliary outcome, the label $y_{\text{out}}^{(1)} \in \{0, 1\}$ indicates post-72-hour in-hospital mortality among patients who remain alive through the 72-hour observation window. This label is intentionally distinct from the terminal reward r_H: the reward reflects survival or death within the logged 72-hour episode, whereas $y_{\text{out}}^{(1)}$ supervises later risk after that window.

More generally, the outcome prediction framework supports endpoints that need not directly correspond to RL rewards:

- **Clinical decompensation:** Deterioration events during the ICU stay
- **Length of stay:** Duration of ICU admission (regression target)
- **Ventilator-free days:** Days alive and free of mechanical ventilation

These auxiliary outcomes can provide additional supervision signal for learning clinically meaningful state representations.

G.2 Why the Auxiliary Head Complements RL

Because the auxiliary outcome targets later risk after the 72-hour window, it complements rather than duplicates the RL reward. It also provides a shorter and less noisy optimization path to the encoder. In a one-dimensional sketch, the supervised gradient is direct:

$$\nabla_\theta \mathcal{L}_{\text{out}} = \frac{\partial \ell}{\partial f_{\text{out}}} \cdot \frac{\partial f_{\text{out}}}{\partial s_H} \cdot \frac{\partial s_H}{\partial \theta},$$

whereas the RL signal must reach the encoder through bootstrapped value estimates and long-

horizon credit assignment:

$$\nabla_\theta J(\pi) \propto \sum_{h=0}^{H_i} \frac{\partial J}{\partial Q_h} \cdot \frac{\partial Q_h}{\partial s_h} \cdot \frac{\partial s_h}{\partial \theta}.$$

Thus the auxiliary head supplies clinically meaningful supervision about longer-horizon risk while giving the representation learner a more direct gradient path with lower variance; RL still determines how that state should be used for sequential control.

G.3 Loss Function

For binary outcomes (post-72-hour mortality, decompensation), we use binary cross-entropy:

$$\ell_k(p, y) = -y \log(p) - (1 - y) \log(1 - p).$$

For continuous outcomes (length of stay), we use mean squared error:

$$\ell_k(p, y) = (p - y)^2.$$

The outcome prediction head is a two-layer MLP:

$$f_{\text{out}}^{(k)}(s_H) = W_2^{(k)} \cdot \text{ReLU}(W_1^{(k)} s_H + b_1^{(k)}) + b_2^{(k)}.$$

In the current experiments, we use $K = 1$ and $\lambda_1 = 1$ for post-72-hour in-hospital mortality after a small search over $\{0.5, 1.0, 2.0\}$, while keeping the general form for extensibility.

H IQL Implementation Details

H.1 Target Network Update

The target value network V_{target} is a slowly-updated copy of V, maintained via Polyak averaging:

$$V_{\text{target}} \leftarrow \tau_{\text{target}} V + (1 - \tau_{\text{target}}) V_{\text{target}},$$

with $\tau_{\text{target}} = 0.005$. This soft update stabilizes training by providing slowly-changing bootstrap targets.

H.2 Expectile Regression

The expectile loss is an asymmetric least squares objective:

$$\begin{aligned} L_\tau^{\text{exp}}(\delta) &= \big(\tau - \mathbb{I}[\delta < 0]\big)^2 \delta^2 \\ &= \begin{cases} \tau\, \delta^2, & \delta \geq 0, \\ (1 - \tau)\, \delta^2, & \delta < 0. \end{cases} \end{aligned}$$

For $\tau > 0.5$, positive residuals (where $Q > V$) receive higher weight, pushing the value function toward the upper portion of the Q-value distribution. This allows IQL to implicitly estimate the value of improved policies without explicitly querying out-of-distribution actions.

The expectile parameter τ controls the trade-off:

- $\tau = 0.5$: Standard symmetric least squares (behavior policy value)
- $\tau \to 1.0$: Approaches maximum Q-value (aggressive improvement)

We use $\tau = 0.7$ as the default, balancing policy improvement with stability.

H.3 Advantage Clipping

For policy extraction, we clip the exponential advantage weights:

$$w_h = \min\big(\exp(A_h/\beta), w_{\max}\big),$$

with $w_{\max} = 20$. This prevents excessively large weights from destabilizing training when the advantage is very positive (indicating actions much better than average).

H.4 Architecture Specifications

Table 23 records the network sizes for the Q, value, and policy heads used in the final IQL stage.

Component	Specification
Q-network	MLP: $d_s + \lvert\mathcal{A}\rvert \to 256 \to 256 \to 1$
V-network	MLP: $d_s \to 256 \to 256 \to 1$
Policy network	MLP: $d_s \to 256 \to 256 \to \lvert\mathcal{A}\rvert$
Activation	ReLU
Output (policy)	Softmax

Table 23: IQL network architecture.

H.5 Hyperparameters

Table 24 lists the retained IQL hyperparameters after the final tuning pass.

Hyperparameter	Value
Expectile τ	0.7
Temperature β	3.0
Weight clipping $w_{\max}$	20
Target update τ_{target}	0.005
Discount γ	0.99

Table 24: IQL hyperparameters.

I Training Details

I.1 Three-Stage Training Procedure

Our training procedure establishes high-quality state representations before policy optimization, preventing *representation drift* where early RL gradients destabilize learned representations.

Stage 1: Representation Pre-training. Train encoder, dynamics model, decoder, and outcome predictor with combined loss:

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{dyn}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{out}},$$

where $\mathcal{L}_{\text{KL}}$ is VAE regularization. This establishes state representations satisfying Q1 (state learning) and Q3 (outcome prediction).

Stage 2: Frozen-Encoder RL. Freeze the encoder (GRU-D, text fusion, VAE posterior) and train only Q-networks, value network, and policy with:

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_Q + \mathcal{L}_V + \mathcal{L}_\pi.$$

Stage 3: Joint Fine-tuning. Unfreeze the encoder with reduced learning rate; auxiliary losses ($\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{out}}$) maintain state quality during adaptation. Table 25 summarizes the three training stages and the corresponding learning rates.

Stage	Epochs	Learning Rate	Components Trained
1: Pre-training	50	1×10^{-3}	Encoder, Dynamics, Decoder, Outcome
2: Frozen RL	100	3×10^{-4}	Q, V, Policy only
3: Fine-tuning	50	1×10^{-4} (encoder) 3×10^{-4} (RL)	All components

Table 25: Training schedule for the three-stage procedure.

I.2 Loss Weights

Table 26 records the fixed Stage-1 loss weights used in the final model.

Loss Component	Weight
Observation reconstruction λ_{obs}	1.0
Mask reconstruction λ_{mask}	0.5
Text reconstruction λ_{text}	0.3
Dynamics loss	1.0
KL regularization	0.1
Outcome prediction	1.0

Table 26: Loss weights for Stage 1 pre-training.

I.3 Entropy Monitoring and Recovery

We monitor policy entropy throughout training to detect collapse:

$$H(\pi) = -\sum_{a \in \mathcal{A}} \pi(a|s) \log \pi(a|s).$$

For a 9-action space, maximum entropy is $\log 9 \approx 2.20$ nats. We flag potential collapse when entropy drops below 0.5 nats or decreases by more than 50% within 10 epochs. Upon detecting collapse, we roll back to the last checkpoint with healthy entropy (> 1.0 nats), reduce learning rate by factor of 2, increase temperature β by factor of 1.5, and resume training.

I.4 Posterior-Collapse Diagnostics

We explicitly monitor posterior-collapse indicators for the latent state. The final model combines three complementary safeguards: KL β-annealing, free-bits, and action-conditioning in the latent dynamics / inference model. Table 27 reports the corresponding latent-usage diagnostics and downstream policy value.

Setting	KL Div.	Active Dims	Mutual Info	FQE
Without mitigation	0.02	3/64	0.17	0.508
β-annealing	4.17	39/64	2.08	0.601
Free-bits	5.74	51/64	2.93	0.619
Full mitigation	8.61	58/64	3.34	0.679

Table 27: Posterior-collapse diagnostics on MIMIC-III. The full training recipe keeps the latent state active and coincides with the best downstream policy value.

Interpretation. The collapsed setting shows near-zero KL divergence and only 3 active latent dimensions. Adding either β-annealing or free-bits improves latent usage, but the strongest non-collapse signal appears only when these are combined with action-conditioning. This is consistent with the view that the belief state is useful because it remains both active and action-sensitive.

I.5 Optimization Details

Table 28 lists the shared optimization settings used across representation learning and policy training.

Setting	Value
Optimizer	AdamW
Weight decay	1×10^{-5}
Batch size	256
Gradient clipping	1.0
Early stopping patience	10 epochs
Validation metric	FQE on validation set

Table 28: Optimization hyperparameters.

I.6 Computational Resources

All experiments were conducted on NVIDIA A6000 GPUs. Training times: Stage 1 ∼4h, Stage 2 ∼5h, Stage 3 ∼2.5h (total ∼11.5h per run). Model size: 1.29M parameters (full model with text fusion).

J Theoretical Analysis

This appendix provides the proof of Theorem 1.

Proof of Theorem 1. Consider the policy gradient for future rewards given current state and action:

$$\frac{\partial}{\partial \pi}\mathbb{E}\left[\sum_{h'=h+1}^{H-1}\gamma^{h'}r_{h'}\;\middle|\;s_h,a_h\right].$$

Under the state decomposition $s_{h'} = g_\theta(\phi_{h'}, z_{h'})$ for $h' > h$, the gradient flows through two pathways:

Pathway 1: Through observations $\phi_{h'}$. In offline RL, observations are fixed in the dataset:

$$\frac{\partial \phi_{h'}}{\partial a_h} = 0 \quad \text{for all } h' > h.$$

Pathway 2: Through latent states $z_{h'}$. Under action-independent dynamics (Definition 1):

$$z_{h+1} = f(z_h, \phi_h, \omega_h) \quad \text{with } \omega_h \perp\!\!\!\perp a_h.$$

Since z_{h+1} does not depend on a_h, and this independence propagates forward:

$$\frac{\partial z_{h'}}{\partial a_h} = 0 \quad \text{for all } h' > h.$$

Combining both pathways via the chain rule:

$$\begin{aligned}\frac{\partial s_{h'}}{\partial a_h} &= \frac{\partial g_\theta}{\partial \phi_{h'}}\cdot\frac{\partial \phi_{h'}}{\partial a_h} + \frac{\partial g_\theta}{\partial z_{h'}}\cdot\frac{\partial z_{h'}}{\partial a_h}\\ &= \frac{\partial g_\theta}{\partial \phi_{h'}}\cdot 0 + \frac{\partial g_\theta}{\partial z_{h'}}\cdot 0\\ &= 0.\end{aligned}$$

Therefore, the reward at any future step $h' > h$ is independent of a_h:

$$\frac{\partial r_{h'}}{\partial a_h} = \frac{\partial R(s_{h'}, a_{h'})}{\partial s_{h'}}\cdot\frac{\partial s_{h'}}{\partial a_h} = 0,$$

and the policy gradient from future rewards vanishes:

$$\frac{\partial}{\partial \pi}\mathbb{E}\left[\sum_{h'=h+1}^{H-1}\gamma^{h'}r_{h'}\;\middle|\;s_h,a_h\right] = 0.$$

□

Implications for Terminal Rewards. When rewards are terminal-only ($r_h = 0$ for $h < H-1$), the policy gradient at any non-terminal step $h < H-1$ depends entirely on future rewards:

$$\nabla_\pi J(\pi)\big|_{s_h} = \mathbb{E}\left[\nabla_\pi \log \pi(a_h|s_h)\cdot Q^\pi(s_h,a_h)\right],$$

where $Q^\pi(s_h,a_h) = \mathbb{E}[\sum_{h'=h}^{H-1}\gamma^{h'-h}r_{h'}|s_h,a_h]$.

Under action-independent dynamics, $Q^\pi(s_h,a_h) = r_h + \gamma\mathbb{E}[V^\pi(s_{h+1})]$ where $V^\pi(s_{h+1})$ is independent of a_h. For $h < H-1$ with $r_h = 0$:

$$Q^\pi(s_h,a_h) = \gamma\mathbb{E}[V^\pi(s_{h+1})] = \text{const w.r.t. } a_h.$$

This means all actions have the same Q-value, making policy improvement impossible.

K Partial Observability Discussion

The latent variable z_h addresses the gap between observable measurements and action-relevant state components that cannot be directly extracted from observations. Examples include:

- **Vasopressor responsiveness:** A patient's responsiveness to vasopressors depends on underlying vascular tone, which is not directly measured but can be inferred from treatment response patterns. Patients with similar blood pressure readings may respond very differently to the same vasopressor dose based on their underlying vascular state.
- **Tissue hypoperfusion severity:** The severity of tissue hypoperfusion may exceed what lactate levels alone indicate, requiring integration of treatment history to estimate. A patient whose lactate remains elevated despite fluid resuscitation may have more severe underlying tissue damage than one with similar lactate levels who has not yet received treatment.

- **Organ reserve capacity:** Organ reserve capacity affects how aggressively a patient can tolerate fluid resuscitation, but manifests only through dynamic responses to interventions. Two patients with similar creatinine levels may have vastly different renal reserves, observable only through how their kidney function responds to fluid challenges.

Through action-conditioned dynamics, z_h can capture how past treatments have shaped the patient's current responsiveness, effectively recovering aspects of the latent health state that are "hidden" in $\tilde{\phi}_h$ but revealed through intervention patterns.

L Evaluation Metrics

L.1 Fitted Q-Evaluation (FQE)

FQE (Le et al., 2019) learns a Q-function for the target policy using the offline dataset:

$$Q^\pi(s,a) \leftarrow r + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s')}[Q^\pi(s',a')].$$

The policy value is estimated as $\hat{V}^\pi = \mathbb{E}_{s_0}[\mathbb{E}_{a\sim\pi(\cdot|s_0)}[Q^\pi(s_0,a)]]$. FQE avoids importance weighting, making it more stable when the learned policy diverges significantly from clinician behavior.

L.2 Weighted Importance Sampling (WIS)

WIS reweights observed trajectories by policy probability ratios:

$$\hat{V}^\pi_{\text{WIS}} = \frac{\sum_{i=1}^N w_i G_i}{\sum_{i=1}^N w_i}, \quad w_i = \prod_{h=0}^{H_i-1} \frac{\pi(a_h^{(i)}|s_h^{(i)})}{\pi_\beta(a_h^{(i)}|s_h^{(i)})},$$

where G_i is the return for trajectory i. WIS is unbiased but has high variance when policies diverge, as indicated by low effective sample size (ESS):

$$\text{ESS} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}.$$

M Off-Policy Evaluation Details

This appendix reports the OPE robustness checks most directly tied to the primary value-estimation results: bootstrap FQE on the standard split, chronological split robustness, and a concrete shadow-mode evaluation note.

M.1 Bootstrap FQE on the Standard Split

Table 29 gives the bootstrap FQE confidence intervals under the same standard split used for the primary result.

Method	MIMIC-III Test FQE	95% CI	Type
Clinician	0.528	[0.520, 0.536]	Behavior
OPL-MT-MNAR (FQE)	**0.679**	**[0.673, 0.686]**	Model-free

Table 29: Bootstrap FQE on the standard random split. The confidence intervals remain separated under the same estimator and held-out protocol.

FQE remains the primary OPE metric here because it is the estimator reported throughout the paper. The CI separation supports a directional value improvement, but it should still be interpreted as observational evidence rather than prospective proof of clinical benefit.

M.2 Chronological Split Robustness

Table 30 checks whether the same policy advantage remains under a temporally ordered train/validation/test split.

The chronological split reduces FQE by 0.014 relative to the standard random split, consistent with mild temporal drift rather than a reversal of the main finding. This helps rule out the possibility that the observed OPE gain is driven by subtle temporal leakage.

M.3 Shadow-Mode Evaluation Note

A feasible prospective protocol is to log AI recommendations in shadow mode, record clinician actions and outcomes, and then analyze concordance, safety events, and counterfactual value estimates under monitoring. The severity-stratified results in Appendix P suggest this is realistic: agreement is naturally higher in low-severity cases and lower where acuity is high and the policy has the most room to differ from standard behavior.

N Baseline Details

This appendix collects the broader baseline families referenced across the paper. Table 31 consolidates the broader sepsis-RL benchmark in one place.

Stronger Missingness-Handling Encoders. Table 32 records the broader MIMIC-III benchmark against stronger sequence encoders for irregular and missing observations. Unlike Table 3, which focuses on a compact prediction comparison, this

Split Strategy	Train	Val	Test	Test FQE	Δ vs Clin.
Random (standard)	–	–	–	0.679 [0.673, 0.686]	+28.6%
Chronological	2001–2008	2009–2010	2011–2012	0.665 [0.656, 0.674]	+25.9%

Table 30: Chronological split robustness on MIMIC-III. The value estimate drops slightly under temporal shift but remains above clinician behavior.

Method	Category	MIMIC-III FQE	AUROC
Continuous State-Space DDQN (Raghu et al., 2017b)	Model-free DDQN	0.476	0.822
Peng et al. 2018	Hybrid (kernel + deep)	0.493	0.828
Model-Based BNN Planner (Raghu et al., 2018)	Model-based BNN	0.498	0.826
DDPG with Clinician Supervision (Huang et al., 2022)	Continuous DDPG	0.529	0.844
AI Clinician	Tabular	0.487	0.812
OPL-MT-MNAR	**MNAR-aware + DocProcess**	**0.679**	**0.886**

Table 31: Broader sepsis-RL benchmark on MIMIC-III. These rows complement the primary baseline table and document comparison to older sepsis-specific baselines.

appendix table makes the broader FQE/AUROC comparison explicit.

Random Policy. Uniform distribution over the 9 actions at each step. Provides a lower bound on policy performance.

Zero-Treatment. Always selects action $(0, 0)$: no fluids and no vasopressors. Tests whether any treatment is beneficial on average.

Clinician (Behavioral Cloning). A policy trained to imitate clinician actions via supervised learning:

$$\mathcal{L}_{\text{BC}} = -\mathbb{E}_{(s,a)\sim\mathcal{D}}[\log \pi_{\text{BC}}(a|s)].$$

Uses the same encoder architecture as OPL-MT-MNAR.

Continuous State-Space DDQN / Model-Based BNN Planner (Raghu et al., 2017b, 2018). These baselines represent the early deep-RL sepsis line: model-free DDQN-style control, continuous-state deep RL, and model-based BNN planning. They are useful historically because they established the 4-hour sepsis RL protocol, but they do not model observation-process MNAR.

Peng et al. (2018) (Peng et al., 2018). A hybrid kernel + deep RL baseline that mixes local similarity structure with learned value estimates. It is a strong pre-offline-RL comparator on MIMIC-III-style sepsis benchmarks.

DDPG with Clinician Supervision (Huang et al., 2022). A continuous-action sepsis RL baseline. It is a direct comparator for the claim that continuous control alone does not solve the MNAR observation problem.

SBCQ (Fatemi et al., 2022). A Semi-MDP / irregular-interval offline RL baseline. It is useful for isolating whether irregular-time handling alone can explain the gains of the proposed observation-process model.

MedDreamer (Xu et al., 2025). A model-based healthcare RL baseline with latent planning. Appendix P additionally reports the high-severity comparison where world-model compounding error is most visible.

Implementation Notes. All methods are compared under their canonical action-space and decision-interval settings in the primary result tables. Cross-setting robustness is reported separately in Table 10 and Table 11, rather than forcing a single action/time discretization onto every baseline.

O Ablation Studies

The controlled studies in this section are grouped by topic for clarity.

O.1 Scope of the Additional Ablations

- **Documentation-process ablations:** Table 18 and Table 20 quantify the contribution of the documentation-process embedding and GRU factor.

Method	Missingness Handling	MIMIC-III FQE	AUROC
Mean Imputation + LSTM	Impute to mean	0.483	0.833
Forward Fill + LSTM	Sample-and-hold	0.488	0.838
GRU-D (Che et al., 2018)	Implicit time decay	0.508	0.844
BRITS (Cao et al., 2018)	Bidirectional imputation	0.516	0.852
mTAND (Shukla and Marlin, 2021)	Multi-time attention	0.523	0.858
MedDreamer	World model + AFI	0.583	0.867
OPL-MT-MNAR	**Explicit MNAR + DocProcess + Text**	**0.679**	**0.886**

Table 32: Appendix benchmark against stronger missingness-handling encoders on MIMIC-III. Explicit MNAR features and documentation-process modeling improve both policy value and representation quality beyond strong irregular-sampling baselines.

- **Temporal granularity controls:** Table 11 reports the 1h / 2h / 4h / 8h decision-interval sweep.
- **Action granularity controls:** Table 10 compares 3×3, 5×5, and continuous action choices.
- **High-frequency text controls:** Table 13, Table 14, and Table 16 show when nursing notes matter most in MIMIC-III.

O.2 Takeaway

Taken together, these controlled studies support the same conclusion as Table 2: the dominant gain comes from preserving observation-process signal, while discretization choice and irregular-time handling act as complementary but secondary factors.

P Additional Subgroup Analysis

Table 33 expands the severity discussion with the explicit low-SOFA versus high-SOFA comparison.

Method	Low SOFA	High SOFA (>10)	Δ
Clinician	0.681	0.192	–
MedDreamer	0.726	0.296	+54.2%
OPL-MT-MNAR	**0.763**	**0.344**	**+79.2%**

Table 33: Severity-focused subgroup comparison on MIMIC-III. The relative advantage is largest in the highest-acuity subgroup, where measurement and documentation intensity are most endogenous.

This subgroup pattern is consistent with the proposed mechanism: observation-process signals are most informative when patients deteriorate and clinicians correspondingly measure and document more aggressively.

Q Clinical Analysis Details

Q.1 Constrained Policy Optimization

To test whether richer state representations remain useful under explicit safety constraints, we add a constrained IQL variant with a Lagrangian penalty:

$$\bar{C} = \frac{1}{H}\sum_{h=1}^{H} C(s_h, a_h),$$

$$\mathcal{L}_{\text{constrained}} = \mathcal{L}_{\text{IQL}} + \lambda\big(\mathbb{E}[\bar{C}] - \kappa\big),$$

with $\kappa = 0.10$. We define $C(s_h, a_h) = 1$ when the action violates rule-based Surviving Sepsis Campaign constraints (e.g., high-dose vasopressors before adequate fluids, or continued aggressive fluids after stabilization). Table 34 compares the unconstrained and constrained variants directly.

Method	Constrained	FQE	Guideline Viol. (%)
Clinician	–	0.528	7.9
IQL (Structured Only)	×	0.591	18.2
IQL (OPL-MT-MNAR)	×	0.679	13.1
C-IQL (Structured Only)	✓	0.582	7.8
C-IQL (OPL-MT-MNAR)	✓	**0.662**	**5.3**

Table 34: Constrained policy optimization on MIMIC-III. Richer MNAR-aware states remain beneficial even after adding explicit guideline constraints.

The constrained variant reduces guideline violations below the clinician baseline while preserving a substantial value advantage, supporting the view that better state estimation and explicit safety constraints are complementary rather than redundant.

Q.2 Text Attention Interpretability

Table 35 lists representative high-attention phrases from the text branch and the clinical contexts they correspond to.

A representative case is the note "worsening hemodynamic instability," which receives high attention on deterioration tokens, a high gate value, and a corresponding increase in vasopressor intensity under the learned policy.

Q.3 Time-to-Deterioration Analysis

We also evaluate the auxiliary head in a time-to-event setting by replacing the binary outcome head

Keyword / Phrase	Attention	Clinical Context
worsening hemodynamic instability	0.231	Cardiovascular deterioration
new onset respiratory distress	0.198	Pulmonary decompensation
increased vasopressor requirements	0.172	Septic shock progression
positive blood cultures	0.143	Infection severity marker

Table 35: Representative high-attention nursing-note phrases (MIMIC-III).

with a discrete-time hazard model over the same 72-hour ICU window. Table 36 summarizes the resulting short-horizon and longer-horizon deterioration metrics on MIMIC-III.

Method	C-index	AOC@12h	AOC@24h	AOC@48h
Cox PH	0.687	0.718	0.703	0.691
Dynamic-DeepHit	0.738	0.769	0.751	0.734
OPL-MT-MNAR (Structured Only)	0.769	0.801	0.781	0.758
OPL-MT-MNAR	**0.821**	**0.859**	**0.832**	**0.801**

Table 36: Time-to-deterioration analysis on MIMIC-III. The MNAR-aware state is especially helpful for short-horizon deterioration prediction, where fresh documentation is most informative.

The largest gain appears at 12 hours, which is consistent with the broader text-MNAR story: documentation-process signals are most useful for near-term changes in acuity.

R Computational Analysis

Model Size. Full model with text fusion: 1.29M parameters. Removing text fusion reduces to 1.05M parameters.

Training Time. On NVIDIA A100 (40GB): Stage 1 pre-training ∼4h, Stage 2 frozen RL ∼5h, Stage 3 fine-tuning ∼2.5h. Total: ∼11.5h per run. Removing text fusion reduces total training time by ∼19% to 9.2h.

Inference. Batch inference (256 patients): 14.2ms latency, 8.4GB memory. This translates to sub-millisecond per-patient decisions, well within real-time clinical requirements.

S Extended Related Work

S.1 Clinical Reinforcement Learning, Offline RL, and Off-Policy Evaluation

Reinforcement learning for critical care has grown substantially since the AI Clinician (Komorowski et al., 2018), which established a common sepsis-RL setup based on 4-hour decision intervals, mortality-related rewards, and off-policy evaluation from observational ICU data. Follow-up work explored continuous state representations (Raghu et al., 2017b), model-based approaches (Raghu et al., 2018), and improved treatment policies under heterogeneous patient responses (Peng et al., 2018; Tang and Wiens, 2021; Huang et al., 2022; Sun and Tang, 2025). More broadly, offline RL has developed methods for stable policy learning under distribution shift, including Batch-Constrained Q-learning (BCQ) (Fujimoto et al., 2019), Conservative Q-Learning (CQL) (Kumar et al., 2020), AWAC (Nair et al., 2021), and Implicit Q-Learning (IQL) (Kostrikov et al., 2022). In parallel, off-policy evaluation (OPE) has provided tools such as importance sampling (IS) (Precup et al., 2000), Per-decision importance sampling (PDIS) (Thomas and Brunskill, 2016), doubly robust estimators (Jiang and Li, 2016), and bootstrap-based uncertainty quantification for assessing policies without deployment (Hanna et al., 2017). In healthcare, these methods are especially important because online exploration is costly or infeasible. Our work adopts this offline RL and OPE perspective, but differs in learning patient states from partial multimodal observations whose missingness patterns are themselves informative.

S.2 Missing Data in Clinical Time Series

Missing data are pervasive in clinical time series and are often MNAR, since measurement and documentation depend on latent patient severity and clinician behavior (Little and Rubin, 2019; Weiskopf and Weng, 2013; Agniel et al., 2018). In structured clinical time series, GRU-D (Che et al., 2018), BRITS (Cao et al., 2018), direct missingness modeling (Lipton et al., 2016), and Raindrop (Zhang et al., 2022b) account for irregular sampling and missingness, but largely focus on prediction rather than sequential decision-making. In multimodal EHR settings, methods such as M3Care (Zhang et al., 2022a), MissModal (Lin and Hu, 2023), DrFuse (Yao et al., 2024), and MUSE (Wu et al., 2024) address missing modalities through fu-

sion, disentanglement, or robustness mechanisms. Our setting differs in two ways. First, we focus on missingness that is endogenously driven by unobserved factors (Xiong and Pelger, 2023; Duan et al., 2024a,b; Li et al., 2024a,b, 2025a; Chen et al., 2026). Second, rather than treating missingness only as a nuisance to accommodate, we use temporally evolving observation patterns across structured data and clinical text as signals for learning patient state. The most closely related work is Liang et al. (2025), which explicitly models informative missingness in multimodal EHR, but without temporal dynamics or policy learning.