MEDICAL WORLD REGULARIZER: LEARNING FROM CLINICAL INTERVENTIONS, HIDDEN STATES, AND FUTURE TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical AI systems predict diverse clinical endpoints from patient observations, but clinical care unfolds in a partially observed world shaped by interventions and revealed through subsequent measurements. A native task specifies what should be predicted; it need not specify which aspects of the evolving patient a representation should retain. **MedWorldReg** assigns longitudinal information three roles: past observations describe a state, prior interventions contextualize a decision, and later observations supervise what an earlier state should preserve. Intervention-aware decision making conditions the native readout on observed care. Future-aware state learning predicts representations of subsequent measurements from an earlier state. The two signals meet in a shared temporal encoder, while the task label, loss, and decision time remain unchanged. This task-world coupling uses no clinical simulator or counterfactual rollout, and its future-prediction branch is absent at deployment. The framework applies to longitudinal tasks with the requisite records, without presuming that every endpoint improves.

1 INTRODUCTION

Medical AI spans diagnosis, patient-state recognition, prognosis, deterioration, and recovery. The observations differ, but the learning contract is often the same: compress information available at a decision point into a representation that predicts a clinical endpoint. This is productive across longitudinal records and imaging (Choi et al., 2016; Li et al., 2020; Elsharief et al., 2025). Yet adding a modality or a longer history need not change what the loss rewards: a feature can help predict the label without describing how the patient reached the observed condition or might evolve from it. Across different clinical tasks, the dominant learning contract remains the prediction of an endpoint.

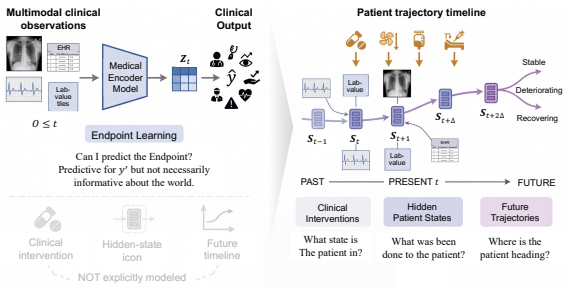


Figure 1: Endpoint learning fits labels; clinical care and future observations reveal an evolving patient.

A patient, however, is not an endpoint. Labs, charts, and images reveal only fragments of physiological condition; when they are obtained is itself informative, and their availability can depend on clinical decisions (Che et al., 2018; Johnson et al., 2023). Clinicians also intervene as the patient evolves. An apparently stable measurement under active support need not imply the same underlying condition as that measurement without support. Later measurements, in turn, carry information about what an earlier representation preserved, although they cannot be consulted when making the earlier decision. These are three distinct roles for observations, care, and subsequent evidence, not three new clinical tasks. Thus *what the task asks the model to predict* differs from *what its representation is asked to preserve*. Endpoint supervision may permit shortcuts without requiring sensitivity to care history or temporal evolution, although it does not necessarily erase either. Figure 1 makes this distinction visible.

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

World models offer a language for observation, hidden state, intervention, and temporal evolution (Ha & Schmidhuber, 2018; Hafner et al., 2025). Recent clinical models use this language to learn patient trajectories or simulate care-dependent transitions (Mu et al., 2026; Wang et al., 2026; Ding et al., 2026), while predictive representation methods learn structure without reconstructing every future measurement (Assran et al., 2023; Adam et al., 2026). These directions reveal a choice often treated as inevitable: keep the endpoint predictor and leave the clinical process implicit, or build a new simulator with its own target and deployment interface. Yet medical records already contain observed care and later measurements alongside native task labels. The question is whether those signals can improve how the *same* task model learns, without assigning them the role of counterfactual outcomes. Can medical AI learn from world structure without having to become a world model?

MedWorldReg answers by assigning each signal a distinct temporal role instead of introducing another clinical task. Observations available at a decision form patient evidence; recorded interventions provide the care context in which that evidence is interpreted; measurements after an earlier training anchor supervise what its state should retain. **Intervention-aware decision making** supplies the native readout with admissible care, and **future-aware state learning** predicts subsequently observed representations from an earlier state. Their task–world coupling occurs in the shared temporal encoder, not in a policy or rollout module. The distinction between the task cutoff and training anchor matters: an observation may be future to the anchor yet already available to a later native decision. This preserves the clinical question and its decision-time information boundary.

The task-preserving view is broader than any one benchmark: diagnosis and prognosis can retain their own labels, metrics, and decision times while using the temporal structure of the records around them. Multimodal benchmarks provide concrete examples of distinct targets under shared data streams (Elsharief et al., 2025; Yin et al., 2026), but they do not establish universal gains. Our claim is instead about *where supervision belongs*: care already delivered can inform a current decision, whereas measurements that arrive later relative to a training anchor can shape its representation without being supplied to that anchor. The resulting hypothesis is testable on each native task rather than requiring the task to be recast as simulation.

Our contributions are:

- **We formulate** endpoint learning against the distinct roles of patient observations, care history, and subsequent measurements.
- **We introduce** intervention-aware decision making and future-aware state learning without changing the native clinical task.
- **We couple** task and temporal supervision in a shared state encoder, without claiming treatment simulation or counterfactual outcomes.

2 Related Work

Endpoint learning from heterogeneous clinical streams. Clinical data arrive as irregular measurements, visits, procedures, and intermittently available modalities. RETAIN and BEHRT model longitudinal visits for future diagnosis, while GRU-D shows that missingness and observation timing themselves can inform prediction (Choi et al., 2016; Li et al., 2020; Che et al., 2018). When records and images are combined, MedFuse studies partially paired streams; MedMod defines several endpoint tasks on linked observations; subsequent benchmark analysis asks when the additional modality helps at all (Hayat et al., 2022; Elsharief et al., 2025; Yin et al., 2026). Temporal generation and alignment go further: DDL-CXR addresses asynchronous imaging, EHRXDiff generates later radiographs from earlier clinical events, and DiPro aligns progression across modalities (Yao et al., 2024; Kyung et al., 2025; Liu et al., 2025). These works establish that clinical streams are heterogeneous and temporally structured; some explicitly model change. The remaining distinction is *supervisory*: MedWorldReg keeps an endpoint task intact while assigning recorded care to its decision context and later observations to the training of an earlier state, rather than asking another modality or another endpoint to carry that burden.

World structure and predictive representations. World-modeling work makes latent states and change under actions explicit (Ha & Schmidhuber, 2018; Hafner et al., 2025). In medicine,

EHRWorld and ChronoMedicalWorld model longitudinal patient trajectories and intervention-conditioned dynamics; CLARITY explores treatment-guided simulation, while medDreamer uses latent imagination for decision support (Mu et al., 2026; Wang et al., 2026; Ding et al., 2026; Xu et al., 2026). Other formulations pursue predictive states without raw-signal reconstruction: I-JEPA and V-JEPA learn to anticipate withheld representations, and SMB-Structure applies latent future prediction to longitudinal records (Assran et al., 2023; Bardes et al., 2024; Adam et al., 2026). Image-world approaches such as CheXWorld and MeWM further show that "world" can denote different transformations and simulation commitments (Yue et al., 2025; Yang et al., 2025). We therefore do not claim latent prediction or intervention modeling in isolation as new. The key shift is to use *world structure as the learning context of a native medical decision*: care conditions its readout, observed futures constrain its shared state encoder, and neither a treatment policy nor a clinical simulator is needed at deployment.

3 Preliminaries: Medical AI as Endpoint Learning

3.1 A unified prediction interface

As illustrated in Figure 1, consider a clinical task with decision time t, label space $\mathcal{Y}$, and native loss ℓ_{task}. Let $o_{\leq t}$ collect all observations available by t, possibly from several modalities and at irregular times. A task model maps these observations to a representation $h_t = B_{\theta,\eta}(o_{\leq t})$ and predicts $\hat{y}_t = H_\phi(h_t)$. The parameters θ denote its temporal component when one exists; η denotes other components, such as imaging. Its native objective is

$$\mathcal{L}_{\text{Med-AI}} = \mathbb{E}\big[\ell_{\text{task}}\big(H_\phi(B_{\theta,\eta}(o_{\leq t})), y_t\big)\big]\,. \tag{1}$$

The notation names an existing objective, not a particular loss or architecture. The label y_t may describe an outcome observed after t, but no observation acquired after t can enter its prediction. The patient's physiological condition s_t is only partly observed through the record; we use s_t conceptually, not as an identified latent recovered by our method. Let $a_{\leq t}$ denote care recorded by the decision time. Clinicians select interventions in response to illness, so their predictive context does not identify treatment effects.

3.2 Task-preserving world learning

The objective in Equation 1 constrains h_t only through its predictive value for y_t. Histories supporting the same endpoint prediction may differ in care context or subsequent trajectory, yet endpoint supervision does not require preserving these distinctions. Temporal supervision provides a complementary constraint on the representation without assuming recovery of the underlying physiological state s_t. Three constraints turn this mismatch into a learning problem. *Task preservation* fixes the native label space $\mathcal{Y}$, decision time t, and loss ℓ_{task}:

$$\mathcal{T}_{\text{Med}} = (\mathcal{Y}, t, \ell_{\text{task}}), \qquad \mathcal{T}_{\text{Med}}^{\text{world}} = \mathcal{T}_{\text{Med}}. \tag{2}$$

Here $\mathcal{T}_{\text{Med}}^{\text{world}}$ denotes the task after adding world regularization: its labels, decision time, and native loss remain unchanged.

Decision-time admissibility allows a task prediction to depend on observations and recorded care only up to t.

Future for learning, not decision uses a separate anchor $\tau < t$: its state cannot see its later training targets. Here t and τ index ordered observation steps after discretization of the irregularly sampled record, and K denotes the future prediction horizon in steps. Writing arrows for information dependence rather than introducing new model mappings, the two cutoffs give

$$\hat{y}_t \leftarrow (o_{\leq t}, a_{\leq t}), \qquad z_\tau \leftarrow o_{\leq \tau}, \quad \text{targets from } o_{\tau+1:\tau+K}. \tag{3}$$

The targets later than τ may still be inside $o_{\leq t}$, the admissible history for the *different* decision at t. No post-t observation enters the task prediction. The method must respect both boundaries while retaining the task in Equation 2.

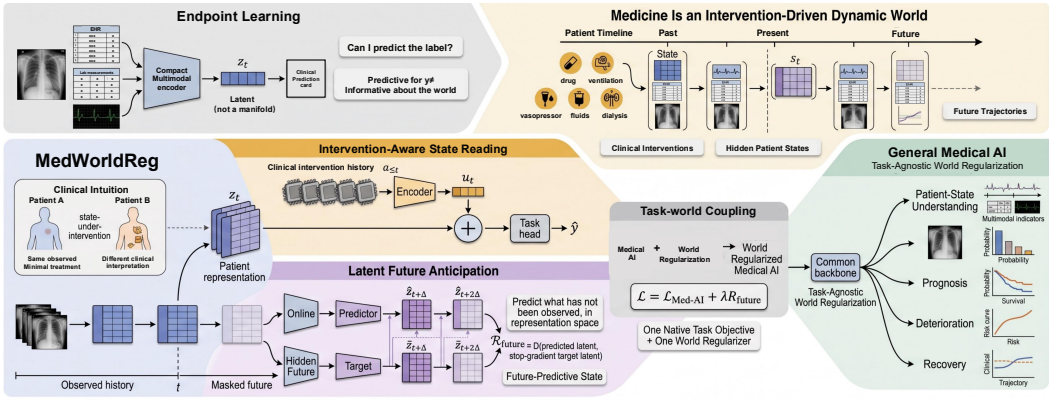


Figure 2: MedWorldReg preserves the native task: care contextualizes its readout, while P_ψ predicts later latent states only during training, without rollout.

4 Medical World Regularizer

Longitudinal records provide three types of evidence: observations describe the patient, actions describe care, and later measurements reveal the subsequent course. Their roles depend on the cutoff. At decision time t, $(o_{\le t}, a_{\le t})$ is admissible prediction context. At an auxiliary anchor $\tau < t$, later observations supervise the representation of $o_{\le \tau}$. We summarize these roles as:

$$\mathcal{W}_{t,\tau} = \big(o_{\le t},\ a_{\le t},\ o_{\tau+1:\tau+K}\big), \qquad \tau < t. \tag{4}$$

The last component is future relative to the anchor, not necessarily to the task decision; its available positions can lie within the task's observed window. $\mathcal{W}_{t,\tau}$ describes temporal information roles rather than a transition model. Observations form patient evidence, interventions provide care context, and subsequent observations supervise an earlier state. The $\mathcal{R}_{\text{future}}$ label in Figure 2 denotes the $\mathcal{R}_{\text{world}}$ objective in Equation 9. **Task supervision says what to predict, and world regularization constrains what the representation should preserve.**

4.1 Intervention-aware decision making

An observed clinical measurement does not have an invariant interpretation independent of the care under which it was recorded. For example, a similar measurement with and without active support can describe different clinical contexts. We therefore form a care summary $u_t = A(a_{\le t})$ and let the native head read a *care-contextualized* representation:

$$h_t = B_{\theta,\eta}(o_{\le t}), \qquad u_t = A(a_{\le t}), \qquad \hat{y}_t = H_\phi(h_t, u_t). \tag{5}$$

The map A makes observed care available to the decision rule without modeling its causal effect. The state h_t can integrate any measurements available by t; the task head reads (h_t, u_t), while the separate future objective acts on a temporal component shared with the state encoder. No particular modality is required by Equation 5.

At the modeling level, the summary exposes three aspects of observed care without imposing a treatment-effect model:

$$u_t = A(a_{\le t}) = \big[u_t^{\text{exposure}};\ u_t^{\text{diversity}};\ u_t^{\text{burden}}\big]. \tag{6}$$

Exposure records which kinds of care have occurred, diversity describes their breadth, and burden summarizes intervention activity. These are interpretable roles, not three new objectives. A can be deterministic; it uses only actions known by the decision time and stays in the deployed readout. It does not directly enter the future predictor, although task gradients can update the shared representation encoder.

4.2 Future-aware state learning

Longitudinal observations also supervise the state without requiring measurement reconstruction. Let $m = (m_1, \ldots, m_T)$ be the time-indexed measurements available in the training record, a tem-

Table 1: Five-task clinical prediction comparison. MedWorldReg is strongest on most reported measures, while general-purpose models lead on remaining-stay estimation. Higher is better except MAD; colored deltas are relative to supervised learning.

	In-hospital mortality		Clinical conditions		Decompensation		Length of stay		Radiology	
Method	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**Kappa**	**MAD**	**AUROC**	**AUPRC**
Native task baselines										
Supervised Learning	81.09	44.32	74.13	42.87	86.67	26.59	37.45	150.1	71.04	31.64
Published methods (adapted)										
Clin-JEPA	81.58 ▲0.49	44.71 ▲0.39	73.62 ▼0.51	41.95 ▼0.92	87.10 ▲0.43	27.08 ▲0.49	37.11 ▼0.34	152.0 ▼1.9	70.16 ▼0.88	30.82 ▼0.82
SMB-Structure	81.32 ▲0.23	43.70 ▼0.62	73.51 ▼0.62	42.16 ▼0.71	86.82 ▲0.15	26.96 ▲0.37	36.91 ▼0.54	149.4 ▲0.7	69.82 ▼1.22	30.92 ▼0.72
SD-JEPA	81.46 ▲0.37	44.03 ▼0.29	73.45 ▼0.68	42.20 ▼0.67	87.04 ▲0.37	27.15 ▲0.56	37.76 ▲0.31	151.6 ▼1.5	69.94 ▼1.10	30.75 ▼0.89
General-purpose models with broad world knowledge										
GPT-5.6	71.62	41.86	66.23	39.07	79.84	25.11	44.83	128.97	64.01	28.89
DeepSeek-V4.1	68.82	39.86	59.25	38.91	76.76	24.35	48.33	124.97	59.13	27.65
Our proposed method										
MedWorldReg	**83.12** ▲2.03	**46.84** ▲2.52	**76.31** ▲2.18	**47.50** ▲4.63	**87.84** ▲1.17	**29.25** ▲2.66	**38.84** ▲1.39	**141.8** ▲8.3	**71.34** ▲0.30	**32.81** ▲1.17

poral component of o; E_θ encodes their history. The state $z_\tau = E_\theta(m_{\leq\tau})$ is *future-aware* when it can predict representations of later observed measurements, not when it identifies physiology or estimates counterfactual outcomes. A withholding operator M_τ preserves the prefix while concealing a later target block in the online view:

$$[M_\tau(m)]_{\leq\tau} = m_{\leq\tau}, \qquad [M_\tau(m)]_{\tau+1:\tau+K} = 0. \tag{7}$$

An EMA target encoder $E_{\bar\theta}$ sees the actually observed, unmasked sequence. The online anchor, target states, and parallel predictions are

$$z_\tau = E_\theta(M_\tau(m))_\tau, \qquad \bar z_{\tau+k} = E_{\bar\theta}(m)_{\tau+k}, \qquad \hat z_{\tau+k} = [P_\psi(z_\tau)]_k. \tag{8}$$

The target is updated by EMA and stopped in the loss; P_ψ maps the single anchor to the entire target block without actions or recursive rollout. The causal encoder already prevents later inputs from reaching z_τ, so the mask makes temporal withholding explicit rather than establishing a new guarantee against a copy predictor.

Let $\mathcal{V}_\tau \subseteq \{1, \ldots, K\}$ be valid positions for an anchor, with $|\mathcal{V}_\tau| > 0$, and let D be a latent-space discrepancy. We regularize the temporal encoder by

$$\mathcal{R}_{\text{world}} = \mathbb{E}_\tau \left[\frac{1}{|\mathcal{V}_\tau|} \sum_{k \in \mathcal{V}_\tau} D\big(\hat z_{\tau+k}, \text{sg}(\bar z_{\tau+k})\big) \right]. \tag{9}$$

Here sg denotes stop-gradient and the expectation is over sampled valid anchors τ. Our instantiation uses latent regression over a bounded subsequent measurement window; the precise horizon and target-update schedule belong to the experimental protocol, while Equation 7 specifies zero masking and the implementation uses SmoothL1 discrepancy. This is neither measurement reconstruction nor simulation. Latent prediction has precedents (Assran et al., 2023; Adam et al., 2026); our contribution is its coupling to the native decision and care context.

4.3 Task–World Coupling

The readout in Equation 5 changes which *admissible* context the task head sees, without introducing another task label or an action loss. Define its native objective under care context and combine it with the single temporal regularizer:

$$\mathcal{L}^{\text{ctx}}_{\text{Med-AI}} = \mathbb{E}\big[\ell_{\text{task}}\big(H_\phi(h_t, u_t), y_t\big)\big]. \tag{10}$$

The native loss still answers the same clinical question. Longitudinal data also contain subsequently observed measurements that can supervise the earlier state at a distinct anchor, as in Equation 9. Assigning each signal its temporal role yields the task-preserving objective:

$$\mathcal{L}_{\text{MedWorldReg}} = \mathcal{L}^{\text{ctx}}_{\text{Med-AI}} + \lambda \mathcal{R}_{\text{world}}, \qquad \lambda \geq 0. \tag{11}$$

Table 2: Ablations of task–world coupling across five clinical tasks. Combining care context with future-state supervision outperforms the isolated components and tested alternative designs.

	In-hospital mortality		Clinical conditions		Decompensation		Length of stay		Radiology	
Method	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**Kappa**	**MAD**	**AUROC**	**AUPRC**
Native task baselines										
Supervised Learning	81.09	44.32	74.13	42.87	86.67	26.59	37.45	150.1	71.04	31.64
Component ablations										
Temporal objective only†	81.72 ▲0.63	45.06 ▲0.74	74.76 ▲0.63	44.18 ▲1.31	87.11 ▲0.44	27.35 ▲0.76	37.98 ▲0.53	147.6 ▲2.5	70.80 ▼0.24	31.54 ▼0.10
Care readout only‡	82.47 ▲1.38	45.97 ▲1.65	75.71 ▲1.58	46.04 ▲3.17	87.39 ▲0.72	28.30 ▲1.71	38.41 ▲0.96	144.6 ▲5.5	71.10 ▲0.06	32.12 ▲0.48
Alternative designs										
Action-Dynamics	79.98 ▼1.11	42.97 ▼1.35	74.46 ▲0.33	42.25 ▼0.62	86.91 ▲0.24	26.83 ▲0.24	36.48 ▼0.97	149.3 ▲0.8	69.70 ▼1.34	30.40 ▼1.24
Raw-Action	81.38 ▲0.29	43.65 ▼0.67	74.41 ▲0.28	43.30 ▲0.43	86.10 ▼0.57	25.74 ▼0.85	36.92 ▼0.53	157.4 ▼7.3	69.96 ▼1.08	31.92 ▲0.28
Bi-Mask JEPA	80.84 ▼0.25	44.57 ▲0.25	73.92 ▼0.21	42.12 ▼0.75	86.94 ▲0.27	26.95 ▲0.36	37.08 ▼0.37	149.2 ▲0.9	70.13 ▼0.91	30.85 ▼0.79
Masked-Value	80.58 ▼0.51	43.52 ▼0.80	74.39 ▲0.26	43.20 ▲0.33	86.18 ▼0.49	25.71 ▼0.88	37.63 ▲0.18	148.9 ▲1.2	70.30 ▼0.74	30.84 ▼0.80
Our Proposed method										
MedWorldReg	**83.12** ▲2.03	**46.84** ▲2.52	**76.31** ▲2.18	**47.50** ▲4.63	**87.84** ▲1.17	**29.25** ▲2.66	**38.84** ▲1.39	**141.8** ▲8.3	**71.34** ▲0.30	**32.81** ▲1.17

† Future-state objective (Equation 9) without the care readout. ‡ Care readout (Equation 5) without the future-state objective.

This is a modular learning principle, not a free performance gain: it uses recorded care as decision context and later observations as training supervision without changing the label, requiring a simulator, or introducing a policy objective. Its value depends on the task and the available longitudinal record. The task readout consumes h_t at cutoff t; the auxiliary objective uses an earlier temporal state z_τ. These need not be identical, but E_θ is shared with $B_{\theta,\eta}$, so both losses shape the same temporal encoder. The target branch is updated by EMA rather than task gradients.

At deployment only $o_{\leq t}$ and $a_{\leq t}$ are required. During training, measurements $\{m_{\tau+k} : k \in \mathcal{V}_\tau\}$ later than the anchor provide auxiliary targets; in the studied setup all their indices satisfy $\tau + k \leq t$. They have a distinct training *role*, not an additional post-t input to the task. The care summary remains in the readout at inference, while P_ψ, masking, and the EMA branch are removed.

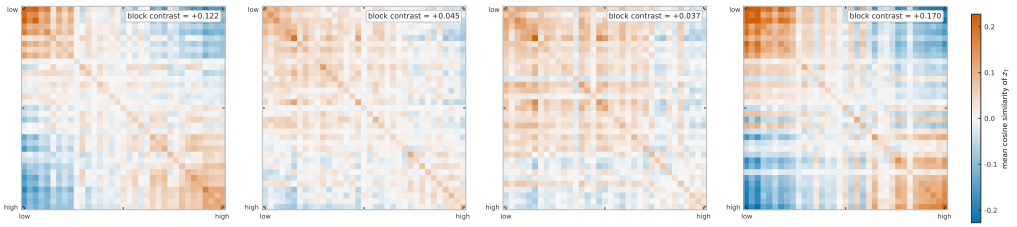
5 Experiments

5.1 Setup

Benchmarks. We evaluate five native decisions on paired ICU histories and radiographs (Johnson et al., 2023; 2019; Elsharief et al., 2025): early in-hospital mortality, end-of-stay multilabel conditions, hourly short-horizon decompensation, hourly remaining length of stay, and radiological findings. They differ in cutoff, endpoint, and observation frequency, but use patient-disjoint partitions. Hourly decisions from one stay are correlated. Detailed task cutoffs, metrics, and implementation scope appear together in Appendix A.

Evaluation protocols. Binary decisions use AUROC and average precision (AUPRC in the tables); multilabel decisions macro-average these over labels. Remaining length of stay uses linearly weighted kappa and the benchmark's reported MAD, for which lower is better. The public MAD routine does not decode ordered classes to duration, so this score should not be read as calibrated error in hours. Colored table deltas are arithmetic differences, not uncertainty estimates. Published ideas are adapted to these native tasks rather than copied from the original papers.

Implementation details. The native predictor encodes longitudinal observations and an available radiograph for a task-specific readout. MedWorldReg adds care context at that readout (Equation 5) and a training-only latent-future objective (Equation 9). Adapted comparisons share an optimizer and learning-rate schedule within this pipeline, though their architecture, original pretraining, and



(a) Supervised (b) Clin-JEPA (c) SD-JEPA (d) MedWorldReg

Figure 3: Final-state similarity ordered by clinical-condition burden. MedWorldReg has the largest within- versus across-burden contrast, revealing a clearer between-patient burden gradient.

compute needs differ. Numerical optimizer settings and adaptation recipes await the completed run record.

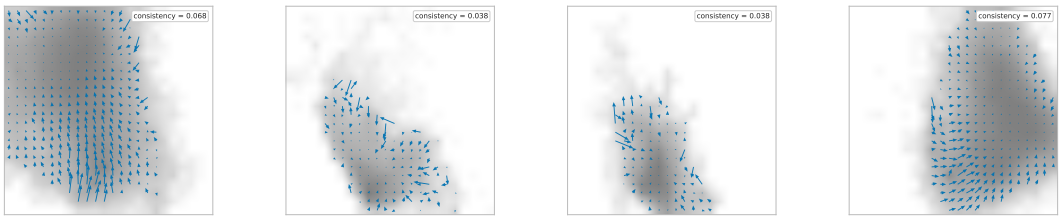
5.2 Main results: a native task benefits from clinical context

Table 1 tests whether the same world-aware principle helps distinct native decisions. **Cross-task benefit without task displacement.** Against supervised learning, MedWorldReg raises mortality AUROC from 81.09 to 83.12 and condition-recognition AUROC from 74.13 to 76.31; hourly decompensation reaches 87.84 AUROC versus 86.67. These tasks ask different clinical questions at different cutoffs, yet all can use the observed sequence of care and subsequent measurements without redefining the label. The gains are consistent with the motivation that an endpoint alone underconstrains how patient state is represented. They do not by themselves isolate which component causes the improvement.

Decision horizon changes the size, not the role, of the gain. For remaining ICU stay, kappa increases from 37.45 to 38.84 while the benchmark's MAD score falls from 150.1 to 141.8; radiology AUROC moves more modestly from 71.04 to 71.34. The contrast is informative: a world-aware representation does not turn every endpoint into an equally dynamic task. Where the target depends on accumulated course, care context and future-aware training have more room to shape the decision; where the image itself strongly defines the label, the improvement can be smaller. This is an interpretation of task structure, not a measured causal decomposition, and the MAD score must not be read as hours saved.

A richer trajectory objective is not automatically a better native decision. Clin-JEPA emphasizes co-trained latent trajectories, SMB-Structure predicts future states from longitudinal records, and SD-JEPA separates progression from content (Yang et al., 2026; Adam et al., 2026; Thil et al., 2026). Their adapted rows improve selected metrics, but none matches MedWorldReg across the completed task families. The comparison points to a specific advantage of our allocation: recorded interventions contextualize the *decision*, whereas later observations train the *state*; predicting more of a trajectory need not provide both. These are adaptations under the present pipeline, not a claim that the original published systems are inferior on their own tasks.

Broad world knowledge is not a substitute for a patient-state representation. The general-purpose models receive structured versions of the same clinical modalities, yet their mortality AUROCs are 71.62 and 68.82, compared with 83.12 for MedWorldReg; on hourly decompensation the corresponding values are 79.84 and 76.76 versus 87.84. The gap is consistent with a distinction between knowing about medicine in general and organizing this patient's noisy, multimodal, hour-by-hour observations and care into a state useful for prediction. The boundary matters: both large models have higher remaining-stay kappa and lower reported MAD than MedWorldReg. Moreover, their scores come from a shared stratified subset rather than the specialized models' full test set (Appendix C); these rows suggest a limitation of general knowledge under the tested interface, not a controlled proof that model scale cannot capture patient dynamics.



(a) Supervised (b) Clin-JEPA (c) SD-JEPA (d) MedWorldReg

Figure 4: Observed latent displacements in a shared projection. MedWorldReg has the highest directional consistency, indicating a more coherent organization of patient-state trajectories.

6 Analysis

6.1 Why both learning principles matter

The two signals answer different questions. In Table 2, the future-state objective of Section 4.2 alone improves mortality and condition recognition but slightly reduces the radiology scores. The care-conditioned readout of Section 4.1 alone improves the displayed scores more broadly. Combining them exceeds either isolated component across the table. This is the distinctive task–world coupling: care makes a current observation interpretable under the support already given, while subsequent observed states constrain what the shared encoder retains. A training-only predictor can therefore alter the deployed representation without being part of the deployed decision. The table supports complementarity among the tested components, not an additive or causal decomposition of their gains.

An unexpected lesson from alternative designs. It might seem more faithful to a world model to feed actions into the transition predictor, or more expressive to use raw action tokens. Yet Action-Dynamics and Raw-Action trail the simpler care-summary readout in Table 2. Bi-Mask JEPA changes temporal withholding and Masked-Value recovers measurements instead of future latents; these also trail the full method, despite selected task-specific gains. The result is not that complexity is intrinsically harmful. It is that the *role* assigned to information matters more than adding it everywhere: observed care is immediately relevant to the native decision, whereas later measurements are useful as a training constraint on an earlier state. Among these tested designs, respecting that division gives the most reliable cross-task pattern.

6.2 The learned state has a direction, not merely a location

Endpoint scores do not show whether a representation organizes patients as evolving states. We inspect three complementary properties: the direction of movement over time, a burden gradient across patients, and changes within one stay. A stay traces a sequence of latent states; the same PCA basis projects all model states for visual comparison, while the reported diagnostics use the full representations. The protocol is in Appendix B. These views ask what structure accompanies better prediction; they are not a substitute for an intervention-conditioned transition test.

Temporal direction: similar positions should imply similar observed change. A collection of endpoint features can cluster patients without encoding which way they are moving. Figure 4 asks a stricter question: among states starting in the same local region, do subsequent observed changes reinforce one another or cancel? MedWorldReg gives the highest consistency, 0.077, versus 0.068 for supervised learning and 0.038 for both adapted JEPA baselines. The advantage aligns with future-aware training: an earlier state is rewarded for anticipating a later one, not merely for explaining a label. The absolute values remain low, so this is increased organization of observed trajectories, not a deterministic transition law or evidence of what would happen under a new treatment.

Between patients: the geometry recovers a clinical burden gradient. Temporal direction alone could encode motion without organizing who is ill. Figure 3 therefore orders patients by an external

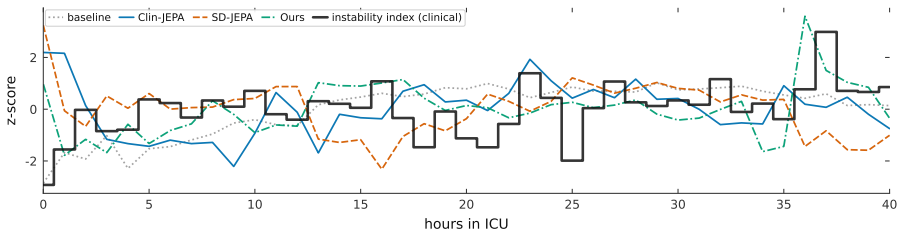


Figure 5: Latent trajectories and vital-sign instability in one ICU stay. MedWorldReg responds to the late rise, illustrating sensitivity to within-patient change in this example.

count of clinical conditions and examines similarity among their final states. The block contrast is highest for MedWorldReg (0.170), compared with 0.122 under supervised learning and 0.045/0.037 for the adapted predictive baselines. This is a different kind of world sensitivity: patients at comparable positions along an observed burden axis occupy more similar regions, while low- and high-burden patients separate. Care-aware reading may help distinguish a measured state maintained by intensive support from one reached without it, and future-aware learning may preserve distinctions relevant to subsequent course. Those are plausible explanations of the geometry, not effects isolated by this sorting test. Burden is a label-derived proxy, not an identified physiological state.

Within a patient: a useful state must move as physiology changes. A patient can be placed correctly among other patients yet still appear static throughout their own stay. Figure 5 addresses this complementary failure mode by plotting latent variation alongside a vital-sign instability index for one stay on the same standardized axis. The MedWorldReg trace responds to some episodes, including a late rise, but also misses others; the displayed trajectory cannot rank models over patients. The patient-wise correlation distribution is not included here. Read together, Figures 4–5 give three distinct questions that endpoint accuracy alone cannot answer: does state have a direction, does it reflect differences in burden between people, and does it register change within a person? The first two receive favorable descriptive evidence here; the third remains illustrative. None of them demonstrates recovered treatment mechanisms, counterfactual prediction, or clinical readiness.

7 Conclusion

Clinical prediction is often trained on endpoints even though a patient's condition evolves under observation and care. MedWorldReg preserves the native prediction task while placing recorded interventions in its decision context and later measurements in a training objective for the earlier state. The two signals meet in a shared temporal encoder; neither a treatment simulator nor a future-prediction branch is needed at deployment. Across five clinical tasks, their coupling improves on isolated components and adapted predictive baselines. The representation analyses suggest that the learned state more clearly organizes observed disease trajectories and between-patient burden, while the general-purpose models' results distinguish broad medical knowledge from tracking this patient's evolving clinical evidence. This is evidence for learning from the structure of the medical world, not for recovering causal treatment mechanisms: the single-stay physiological trace is illustrative, and large-model estimates use a sampled test subset. The remaining-stay result also shows that task-specific strengths are not universal. More stringent temporal audits, patient-level uncertainty, and external validation are required before this representation can support clinical decisions.

AI use statement

Generative AI was used to help formulate and critique the paper's conceptual framing, draft mathematical exposition and manuscript text, identify and format literature references, inspect figures and supplied aggregate result tables, and assist with figure layout. It was not used here to generate experimental measurements, analyze patient-level data, or implement a new clinical model. The authors must independently verify the cited sources, numerical claims, run provenance, equations, figures, and all AI-assisted prose before submission; they remain responsible for the final content and for disclosing any additional AI use in the completed study.

Ethics statement

This work concerns retrospective clinical prediction, not treatment recommendation or causal effect estimation. Recorded interventions can reflect both disease severity and care decisions, and observational associations should not be interpreted as evidence of therapeutic benefit. Any deployment would require evaluation for subgroup performance, calibration, temporal leakage, and shifts in data collection and clinical practice.

Reproducibility statement

The native task cutoffs, label definitions, and evaluation measures are summarized in Section 5.1 and Appendix A. Equations 5–11 specify the care-context readout and temporal objective; Appendix B defines the representation diagnostics. Exact run configurations, adapted-method recipes, patient-level uncertainty, and the hourly image-timestamp audit must be supplied before final submission. We do not present the current manuscript as a complete reproducibility package.

References

Irsyad Adam, Zekai Chen, David Laprade, Shaun Porwal, David Laub, Erik Reinertsen, Arda Pekis, and Kevin Brown. The Patient is not a Moving Document: A World Model Training Paradigm for Longitudinal EHR, 2026. URL https://arxiv.org/abs/2601.22128.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, 2016. URL https://proceedings.neurips.cc/paper/2016/file/231141b34c82aa95e48810a9d1b33a79-Paper.pdf.

Tianxingjian Ding, Yuanhao Zou, Chen Chen, Mubarak Shah, and Yu Tian. CLARITY: Medical World Model for Guiding Treatment Decisions by Simulating Context-Aware Disease Trajectories. In *Computer Vision – ECCV 2026*, pp. 202–221, 2026. doi: 10.1007/978-3-032-37038-9_12. URL https://doi.org/10.1007/978-3-032-37038-9_12.

Shaza Elsharief, Saeed Shurrab, Baraa Al Jorf, Leopoldo Julian Lechuga Lopez, Krzysztof J. Geras, and Farah E. Shamout. MedMod: Multimodal Benchmark for Medical Prediction Tasks with Electronic Health Records and Chest X-Ray Scans. In *Proceedings of the Sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pp. 781–803. PMLR, 2025. URL https://proceedings.mlr.press/v287/elsharief25a.html.

David Ha and Jürgen Schmidhuber. World models, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nasir Hayat, Krzysztof J. Geras, and Farah E. Shamout. MedFuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pp. 479–503. PMLR, 2022. URL https://proceedings.mlr.press/v182/hayat22a.html.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0.

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

Daeun Kyung, Junu Kim, Tackeun Kim, and Edward Choi. Towards Predicting Temporal Changes in a Patient's Chest X-ray Images based on Electronic Health Records. In *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pp. 247–267. PMLR, 2025. URL https://proceedings.mlr.press/v287/kyung25a.html.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10:7155, 2020. doi: 10.1038/s41598-020-62922-y.

Chen Liu, Wenfang Yao, Kejing Yin, William K. Cheung, and Jing Qin. Multimodal Disease Progression Modeling via Spatiotemporal Disentanglement and Multiscale Alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, Main Conference, pp. 132539–132566. Curran Associates, Inc., 2025. doi: 10.52202/085713-4415. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/bfe27c959fda8112520479539c69360f-Paper-Conference.pdf.

Linjie Mu, Zhongzhen Huang, Yannian Gu, Shengqian Qin, Shaoting Zhang, and Xiaofan Zhang. EHRWorld: A Patient-Centric Medical World Model for Long-Horizon Clinical Trajectories, 2026. URL https://arxiv.org/abs/2602.03569.

Lucas Thil, Jesse Read, Rim Kaddah, and Guillaume Doquet. Subspace-decomposed jepas: Disentangling progression and content in latent world models, 2026. URL https://arxiv.org/abs/2605.31111.

Jiangyuan Wang, Xuyong Chen, Junwei He, Xu Xu, Shasha Xie, and Fuman Han. ChronoMedicalWorld: A Medical World Model for Learning Patient Trajectories from Longitudinal Care Data, 2026. URL https://arxiv.org/abs/2605.21963.

Qianyi Xu, Gousia Habib, Feng Wu, Dilruk Perera, and Mengling Feng. medDreamer: Model-Based Reinforcement Learning with Latent Imagination on Complex EHRs for Clinical Decision Support. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 1693–1704, 2026. doi: 10.1145/3770854.3780292. URL https://doi.org/10.1145/3770854.3780292.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, and Jieneng Chen. Medical World Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8319–8329, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_Medical_World_Model_ICCV_2025_paper.html.

Yixuan Yang, Mehak Arora, Ryan Zhang, Baraa Abed, Junseob Kim, Tilendra Choudhary, Md Hassanuzzaman, Kevin Zhu, Ayman Ali, Chengkun Yang, Alasdair Edward Gent, Victor Moas, and Rishikesan Kamaleswaran. Clin-JEPA: A multi-phase co-training framework for joint-embedding predictive pretraining on ehr patient trajectories, 2026. URL https://arxiv.org/abs/2605.10840.

Wenfang Yao, Chen Liu, Kejing Yin, William K. Cheung, and Jing Qin. Addressing Asynchronicity in Clinical Multimodal Fusion via Individualized Chest X-ray Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 29001–29028. Curran Associates, Inc., 2024. doi: 10.52202/079017-0913. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3310034c97fab48fdbcba18f90fd5364-Paper-Conference.pdf.

Kejing Yin, Haizhou Xu, Wenfang Yao, Chen Liu, Zijie Chen, Yui Haang Cheung, William K. Cheung, and Jing Qin. When Does Multimodal Learning Help in Healthcare? A Benchmark on EHR and Chest X-Ray Fusion, 2026. URL https://arxiv.org/abs/2602.23614.

Yang Yue, Yulin Wang, Chenxin Tao, Pan Liu, Shiji Song, and Gao Huang. CheXWorld: Exploring Image World Modeling for Radiograph Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20778–20788, 2025. doi: 10.1109/CVPR52734.2025.01935. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yue_CheXWorld_Exploring_Image_World_Modeling_for_Radiograph_Representation_Learning_CVPR_2025_paper.html.

A Task and Evaluation Protocols

A.1 Paired cohort and admissible observations

The benchmark links ICU observations in MIMIC-IV to chest radiographs in MIMIC-CXR within the same stay (Johnson et al., 2023; 2019; Elsharief et al., 2025). Eligible images are frontal AP views; stays without a usable EHR–radiograph pair do not enter the paired cohort. The clinical stream comprises 17 groups of bedside and laboratory measurements, including vital signs, Glasgow Coma Scale components, glucose, and blood pH. Its gridded input retains observed-value indicators alongside imputed measurements, producing a 76-dimensional vector per time step. Missing values are carried forward when a previous measurement exists; otherwise the benchmark supplies predefined defaults. The available documentation does not resolve whether the runs reported here used one- or two-hour gridding, so no specific sampling interval is assumed below.

Patients are partitioned across training, validation, and test sets rather than splitting individual hourly examples. The documented target allocation is approximately 70/10/20 percent by patient; the numbers of instances differ by task because a stay may yield multiple prediction times or multiple radiographs. The prediction cutoff is part of each task definition, not a property shared by all five tasks. In particular, a dynamic decision should use only measurements, images, and recorded care already available at its cutoff. The documented image-pairing process does not establish that this condition was enforced independently at every hourly cutoff; a strict no-future-image claim awaits verification of the actual pairing and timestamp logic.

A.2 Five native prediction tasks

Early in-hospital mortality. The input window is the first 48 hours following ICU admission, together with the latest eligible radiograph acquired in that window. The binary target is death during the hospitalization, including death after the 48-hour observation window. This is an early-risk task, not a prediction of death within the first 48 hours.

Clinical conditions. At the end of the ICU stay, the model assigns 25 non-exclusive condition labels using the longitudinal clinical record and the latest eligible in-stay radiograph. The conditions cover acute and chronic diagnoses, including renal, cardiovascular, pulmonary, infectious, and metabolic conditions. Because both observations and labels extend through the stay, this experiment must not be interpreted as admission-time prospective diagnosis.

Hourly decompensation. Beginning four hours after ICU admission, each eligible hour provides a decision point using the clinical history through that hour. The binary target indicates whether death occurs within the following 24 hours; the documented labeling convention treats the exact

24-hour boundary as negative. An image used for an hourly decision would have to precede that decision, subject to the timestamp audit noted above. Multiple hours from one stay are correlated, not independent patient observations.

Hourly remaining ICU stay. This task also begins four hours after admission and predicts the time from the current hour to ICU discharge. The original remaining duration is assigned to ten ordered bins: less than one day; successive one-day bins through seven to less than eight days; eight to less than 14 days; and 14 days or more. As with decompensation, both the clinical sequence and any paired image should be restricted to information available at the decision hour, and individual hours must not be treated as distinct patients.

Radiological findings. An eligible chest radiograph from an ICU stay supplies an image-specific 14-label target derived from its report, with the associated clinical record used as context. A stay can contribute more than one examination; the documented reader uses one eligible image record per examination and clinical history up to the image time for this task. The radiology report supplies labels rather than an input modality. This is therefore not a claim that the model reads the entire subsequent hospital course for every image.

A.3 Metrics, aggregation, and interpretation

In-hospital mortality and hourly decompensation use AUROC and average precision. For clinical conditions and radiological findings, these are computed per label and macro-averaged across labels; the "AUPRC" heading in the result table denotes average precision, not trapezoidal area under the precision–recall curve. The benchmark evaluation treats uncertain or unmentioned radiological findings as negative. For the ten ordered remaining-stay classes, linearly weighted Cohen's κ penalizes predictions more as they depart further from the true bin. The reported MAD field inherits the public benchmark routine, which computes absolute error using an output class without first decoding it to a remaining duration. We therefore do not interpret that field as calibrated error in hours or as median absolute deviation. Table deltas are arithmetic differences from the supervised reference, not confidence intervals or significance tests. Uncertainty over hourly examples should respect patient clustering rather than resampling correlated hours as independent patients.

A.4 Model instantiation and reproducibility

The evaluated instantiation combines a temporal clinical encoder and an image encoder for each native task. At decision time, Equation 6 has three parts: $u_t^{\text{exposure}} \in \mathbb{R}^8$ records a presence indicator for each of eight action groups (vasopressors, ventilation, dialysis, sedation, antibiotics, transfusion, fluids, and laboratory orders); $u_t^{\text{diversity}} \in \mathbb{R}^8$ records the number of distinct action types per group divided by observation-window duration; and $u_t^{\text{burden}} \in \mathbb{R}^2$ holds the total number of distinct action types and its per-hour rate. These count distinct types, not repeated administrations. Action features are standardized using the training partition and restricted to the observation window. During training, a causal online encoder predicts latent targets in a contiguous future block from an earlier anchor; an EMA target encoder sees the unmasked sequence and provides stopped-gradient targets. A SmoothL1 objective trains the parallel predictions. No action-conditioned rollout is used at inference. Exact optimizer settings, learning-rate schedule, adaptation recipes, random seeds, and run-level uncertainty must be supplied from the completed experiment record rather than borrowed from unrelated pretraining protocols.

B Representation Diagnostics

For each stay, let $\tilde{z}_t$ be the temporal encoder's representation at hour t after a frozen diagnostic projection head. This differs from the online anchor z_τ in Equation 8; the projection head is used only to compare representations, not as a task input or future predictor. The last observed state $\tilde{z}_T$ summarizes the stay. A single PCA basis is fit jointly to the resulting states from all compared methods and used to project every panel; the same axes and arrow scaling are applied throughout. This common projection supports side-by-side visual comparison but does not preserve all latent-space distances. Scalar diagnostics below are computed before PCA projection unless otherwise stated.

Temporal direction (Figure 4). Let $\Delta\tilde{z}_t = \tilde{z}_{t+1} - \tilde{z}_t$ for adjacent valid hours. Project starting states and displacements onto the shared PCA plane and divide it into a 24×24 grid. In each cell with at least 25 starting transitions, draw the mean projected displacement using one global arrow-length scale; leave sparser cells empty. The scalar label is the full-dimensional ratio $\|\mathbb{E}[\Delta\tilde{z}]\|_2/\mathbb{E}[\|\Delta\tilde{z}\|_2]$, bounded between zero and one. Equal-sized displacements pointing the same way approach one, whereas opposing directions cancel in the numerator. It describes coherence among observed latent transitions, not whether treatment produces an identified state transition or whether any specific patient can be forecast.

Between-patient burden (Figure 3). For each stay, use the final projected-head state $\tilde{z}_T$. Center these vectors across the cohort and compute pairwise cosine similarities; ordering is defined by an external burden measure, the number of positive labels among the 25 clinical conditions. Partition the ordered patients into equal-sized bins and display mean similarity between bins, with the same color scale for every method. Define block contrast as half the sum of within-low-tercile and within-high-tercile similarity, minus similarity across those terciles. The score captures separation in this label-derived ordering, not the identifiability of an unobserved severity variable. The ordering and grouping are analysis operations, never model inputs. Any later report should specify the number of bins, handling of ties, and diagonal/self-similarity convention.

Within-patient change (Figure 5). The reference instability index uses observed heart and respiratory rates with positive signs and oxygen saturation and systolic blood pressure with negative signs. Standardize each vital sign over the analysis cohort first; at a given hour average the signed standardized values only if at least three of four are actually observed. Larger values indicate greater vital-sign instability. Within each stay, standardize the clinical index and the first latent principal-component trajectory separately before putting them on the same vertical scale. Figure 5 displays only one stay. For a cohort-level statistic, compute Spearman correlation within each eligible patient between their PC1 trajectory and instability index, requiring at least ten valid hours and a nonconstant index; report the distribution across patients, not merely the mean. That distribution has not been included in the current figure and must not be inferred from this example. PCA sign is arbitrary; any orientation convention for comparing correlations across methods must be fixed and reported before the patient-level analysis.

C Structured Inputs for General-Purpose Models

C.1 Paired zero-shot evaluation

The general-purpose multimodal models in Table 1 draw from the same paired held-out ICU test pool and native task definitions as the specialized-model evaluation. For each selected example, the native observation and care inputs are converted into one structured text message, with its eligible radiograph supplied separately as an image. No patient outcome, radiology report, future observation relative to the task cutoff, or worked answer is included in the prompt. Each example is an independent zero-shot request; all compared general-purpose models receive the same selected examples and task template. Unlike the full test-set evaluation, these requests cover a sampled subset; direct paired comparison to a specialized model requires scoring that model on the same subset.

For each task the protocol selects 250 test instances by fixed stratified sampling, shared across general-purpose models. Mortality is stratified by its binary label at the cohort rate, while the two multilabel tasks are stratified by the number of positive labels. Decompensation balances its positive and negative classes; remaining stay samples equally across its ten ordered target bins. Population estimates account for the stratum inclusion probabilities, particularly for prevalence-sensitive metrics and weighted κ; the analysis must preserve the patient-level clustering of hourly examples. These sampling and weighting choices belong to the evaluation protocol, not to the prompt; they do not reveal labels to the queried model.

C.2 From clinical tensors to readable histories

The clinical input is derived from the same preprocessed observation window rather than from a new review of the raw record. Continuous channels are restored from normalized values to their clinical

scale, and categorical channels are decoded from their one-hot representation. Observation indicators distinguish a value measured in the current time bin from a carried-forward value (marked ~) or a default used before the first measurement (marked ^). Each channel has a full-window summary (observed count, range, first and last values for continuous channels; observed count, mode, and modal fraction for categorical channels), followed by the last 24 hours of time-resolved values. A channel never measured in the window is identified as such rather than repeatedly listed. This compact rendering retains the clinical values and missingness distinctions supplied by preprocessing, but the full-window summary is a compression of older time steps and should not be described as a lossless textual copy of the complete tensor.

The care block renders the same action summary available to the native readout. For each of the eight care groups in Section A.4, it gives a presence indicator and the number of *distinct* action types per observation-window hour; it also gives the total number of distinct types and its per-hour rate. These are type counts, not counts of administrations, and a rate is not a percentage. The radiograph is the selected image from the paired example without training-time augmentation; the text states its acquisition time relative to ICU admission. For an hourly decision, eligibility of that image must still be checked at the prediction cutoff, as noted in Section A.1.

C.3 Task prompts and response format

The following schematic shows the system instruction and the shared user-message structure. Bracketed slots describe fields filled separately for each example; they are not patient data or demonstration answers. The complete condition/finding vocabulary is supplied in the corresponding task block, in the same order for every model.

```
SYSTEM: You are an expert critical-care physician.  Answer
one prediction task from the supplied ICU data and attached
chest X-ray.  Do not invent measurements.  Return only the
requested JSON.
USER: ICU stay [anonymous identifier]
Chest X-ray acquired [hours] after ICU admission:  [image]
EHR: [window length and time bin width]; for each channel,
[whole-window summary] and [recent time-resolved values].
A bare value was measured in that bin; ~ means carried
forward; ^ means default before any measurement.
ACTIONS: for each care group, [distinct types/hour,
present]; [total distinct types, total types/hour].
TASK: [task-specific prediction target, label list when
applicable, and response schema].
```

The task block and required JSON keys vary only with the native target:

- **In-hospital mortality:** Predict death during hospitalization from the early window; return `{"present": Boolean, "confidence": [0,1]}`.
- **Hourly decompensation:** Predict death within the next 24 hours; return the same binary JSON keys.
- **Clinical conditions:** Judge each of the 25 named conditions independently at the end of the stay; return `{"labels": {"<condition>": {"present": Boolean, "confidence": [0,1]}}}` for every condition.
- **Radiological findings:** Judge each of the 14 named findings for the attached examination; return the same per-label JSON structure for every finding.
- **Remaining ICU stay:** Predict the remaining duration and one of ten ordered bins; return `{"predicted_remaining_hours": number, "predicted_bin": integer, "confidence": [0,1]}`.

For AUROC and average precision, the positive-class score is the reported confidence when `present` is true and one minus that confidence otherwise; weighted κ uses the predicted ordinal bin. Predicted hours are not substituted for a decoded bin in the benchmark's MAD calculation described in Section A.3.