SIGN-CONSTRAINED INTERVENTION EFFECTS FOR DOMAIN-GENERALIZABLE ICU WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting how a patient's vital signs will respond to an intervention is a central question in intensive care. World models can do so by learning dynamics as a function of prior actions, so that a trajectory can be unrolled under any chosen treatment course. However, such models tend to be brittle for dosing practices outside of those seen in training data. Clinicians choose drug dosages based on the patient's state, so the association a model learns between dose and outcome runs opposite to the drug's *effect*, which does not generalize well to out-of-distribution (OOD) dosing practices. However, pharmacological information already provides the effects on vital signs of different drugs, which is invariant to dosing practice distribution shifts. To resolve the OOD sensitivity issues of current ICU world models, we introduce PHYSIOWORLD, a world model that isolates an intervention's effect by evaluating a forward pass twice: once under the recorded doses, and once with those doses set to zero. The difference is then projected onto the set of pharmacologically admissible *effect directions* stated by a corpus of 70 (drug, channel, sign) rules over 28 drugs, leaving the magnitude, which no source states, to be learned from data. Across five distribution-shift parameters over cohorts from three intensive-care databases, and against forecasters, counterfactual treatment-effect models, invariance objectives and two weaker placements of the same knowledge, PHYSIOWORLD attains the lowest OOD RMSE in every setting, improving on the strongest external baseline by 8–12%, while matching the in-distribution error of an unconstrained backbone. Reversing the pharmacological effect directions degrades accuracy below the unconstrained backbone, and directions estimated from the training records recover almost no performance gain, indicating that the improvement derives from the content of the pharmacological knowledge rather than from the presence of a constraint. The construction is not tied to intensive care and may extend to other settings where the sign of an effect is known in advance and its magnitude is confounded by how the treatment was assigned.

1 INTRODUCTION

In an intensive care unit, clinicians must frequently choose interventions to perform on a patient, such as raising a vasopressor, starting a sedative, or changing a ventilator setting. Each decision rests on an implicit prediction of how the patient's vital signs will respond, and making that prediction explicit is a central computational problem in this setting (Hyland et al., 2020; Komorowski et al., 2018). The data is two streams on one clock: vital signs sampled by clinical routine as much as by physiology (Agniel et al., 2018; Hinrichs et al., 2024; Harutyunyan et al., 2019), and sparse, asynchronous clinician actions that mix drug infusions, ventilator set-points and discrete switches. Clinical time-series models handle such data natively (Che et al., 2018; Rubanova et al., 2019; Nie et al., 2022) but answer only "what happens next", whereas the question here is what happens under a treatment course one names, which requires interventions to be explicit inputs whose consequences propagate through the predicted physiology, the defining property of a *world model* (Ha & Schmidhuber, 2018; LeCun et al., 2022).

The difficulty is not in fitting such a model, but in applying it to units, eras, or dosing practices unlike those of the records it was trained on, which is where it is used and where every data-driven one fails. That failure is documented: on sepsis cohorts from MIMIC-IV and eICU, Park et al. (2024) trained

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.

81 transformer dynamics models and found that replacing the recorded doses at test time with zeros or with a random permutation left prediction error essentially unchanged. What they carry is not a dose-response, and a model that holds none has nothing that a change of regime could preserve.

The cause is in the data rather than the architectures. A clinical record contains only the treatments that were given, and the dose is itself a reading of how sick the patient was when it was set (Gottesman et al., 2019; Kotani et al., 2023). A vasopressor appears in the record because the patient was deteriorating, so a model that reads the dose as a cause rather than as a symptom of severity learns that the drug makes the pressure fall (Walker, 1996; Hernán & Robins, 2010; Ioannidis et al., 2001). That accounts for both symptoms: where dose and severity are collinear the fitted effect collapses toward zero, and where severity dominates it takes the wrong sign. Either way it surfaces out of distribution, since a dose-response learned this way survives only the regime it was fitted in.

The failure concentrates on one axis: every purely data-driven model we test degrades further out of distribution under a shift in treatment intensity than under any shift in the patient population (Section 5.2). A model whose dose-response is absent or reversed has nothing to carry across a change in dosing, while the rest of the physiology it learned survives intact. The *direction* of a drug's effect, by contrast, is invariant across that shift: magnitudes vary with age, organ function and comorbidity, but a vasopressor raises blood pressure across patients and care units, and pharmacology states as much before any data is collected (Hall & Hall, 2021).

PHYSIOWORLD turns that asymmetry into an architecture. A standard forward pass does not expose the quantity a direction would constrain, since it outputs the next state as a whole, without separating the part caused by any single drug. PHYSIOWORLD recovers that part by evaluating the same predictor twice with identical weights and history, once with the recorded doses and once with those doses set to zero, and a closed-form, parameter-free projection then maps the difference onto the cone of pharmacologically admissible directions. The operator therefore constrains signs alone and leaves magnitudes to training (Section 4). We evaluate PHYSIOWORLD on five distribution-shift axes across three ICU databases, where it exceeds every baseline out of distribution at the in-distribution error of its backbone, and the gain holds under free-running rollout (Section 5).

This paper makes three contributions.

- We build a shift benchmark on five cohorts from three intensive-care databases, spanning patient population and treatment intensity, and show that clinical dynamics models degrade further under a shift in how much drug is given than under any shift in who the patient is.
- We close that gap with PHYSIOWORLD, a clinical world model that isolates its estimate of an intervention's effect with a second forward pass and projects that estimate onto the stated pharmacological directions, a parameter-free, provably non-expansive operation that reduces out-of-distribution RMSE by up to 12% over the strongest external baseline.
- We ablate the injected corpus out of distribution. Inverting the stated directions is worse than injecting none, and data-mined directions recover little of the benefit, so the generalization gain comes from the content of the pharmacology rather than from the act of constraining.

2 RELATED WORK

World models. Action-conditioned latent dynamics were first developed in games and robotics (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2025), and joint-embedding predictive architectures (JEPA) later showed that predicting in representation space, without reconstructing the input, is sufficient for the same purpose (LeCun et al., 2022; Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025; Balestriero & LeCun, 2025; Bagatella et al., 2026; Chen et al., 2026), a recipe our backbone follows. Several world models have recently been proposed for medicine (Liu et al., 2026a; Yang et al., 2026; 2025; Wang et al., 2026; Mu et al., 2026; Liu et al., 2026b), but none of the six applies to the present task: four take no intervention input at all, one models an intervention as producing an empty outcome with probability one, and one encodes vasopressor exposure as a binary flag, so a treatment course cannot be specified. None of the six builds pharmacological structure into the architecture, and none is evaluated under distribution shift.

Clinical prediction, interventions and shift. Models for bedside data address irregular sampling directly (Che et al., 2018; De Brouwer et al., 2019; Rubanova et al., 2019; Yalavarthi et al., 2024), while windowed and pretrained forecasters remain strong on aggregate error (Zeng et al., 2023; Nie

et al., 2022; Liu et al., 2024; Wang et al., 2024; Ansari et al., 2025; Li et al., 2026; Gu et al., 2024; Pucher et al., 2026; Meyer et al., 2025), and in all of them interventions enter as covariates. A second line estimates counterfactual outcomes under treatment that changes over time, adjusting for time-dependent confounding with balanced representations (Bica et al., 2020) or attention over treatment history (Melnychuk et al., 2022); both are baselines here, and both face the confounding described above (Walker, 1996; Hernán & Robins, 2010; Gottesman et al., 2019), which is why only knowledge that does not require estimating response magnitudes from the record can be injected safely. Robustness to shift is otherwise sought through invariant features (Arjovsky et al., 2019; Sun & Saenko, 2016; Sagawa et al., 2019), and clinical shift benchmarks stratify patients by demographics, care unit or era (Avati et al., 2021; Koh et al., 2021; Yang et al., 2023), all of which describe who the patient is.

Knowledge-constrained learning. A common approach relaxes logic or physics constraints into a differentiable penalty (Xu et al., 2018; Raissi et al., 2019; Badreddine et al., 2022; Huang & Raza, 2026), which leaves the model free to violate the rules it was penalized with. A second approach enforces them exactly by projecting the output onto a feasible set (Amos & Kolter, 2017; Donti et al., 2021; Dalal et al., 2018). A third builds the constraint into the function class, as in monotonic networks (Sill, 1997; You et al., 2017; Liu et al., 2020) and physics-informed architectures (Greydanus et al., 2019; Cranmer et al., 2020), and pharmacological priors have entered clinical state-space models in a similar way (Hussain et al., 2021; Qian et al., 2021). All of these constrain a predicted *value*, or its derivative with respect to a designated input, and none reaches the quantity at issue: an intervention effect can contradict the direction of a drug's action while every value it produces stays feasible, because monotonicity in a raw input places no constraint on a dose effect that has already propagated through the latent representation, as our Clamp baseline shows. We therefore place the constraint on Δz itself, inside the recurrent transition.

3 Problem Setup

Problem formalization. A patient stay is a triple of streams on a common clock (Figure 1). The *state stream* is a sequence $\{(\tau_t, x_t, m_t)\}_{t=1}^{T}$ of irregularly spaced times $\tau_1 < \cdots < \tau_T$, last-observed vital-sign vectors $x_t \in \mathbb{R}^C$, and freshness masks $m_t \in \{0,1\}^C$, where $m_{t,c} = 1$ if and only if channel c is newly measured at τ_t. Between measurements x carries the last value forward, so the mask, not the value, encodes what is known. The *intervention stream* is a set of typed events $\mathcal{E} = \{(s_i, k_i, v_i)\}$: administration time s_i in continuous hours, intervention channel $k_i \in \{1, \ldots, A\}$ (drug infusions and boluses, ventilator set-points, discrete switches), and dose or setting v_i. The third stream is the static covariate vector ξ (e.g. age, sex, care unit), fixed for the stay. The task is one-step state prediction under a specified intervention course: given the history $\mathcal{H}_t$ of all three streams up to τ_t and the interventions in $(\tau_t, \tau_{t+1}]$, predict x_{t+1} at the next observed time,

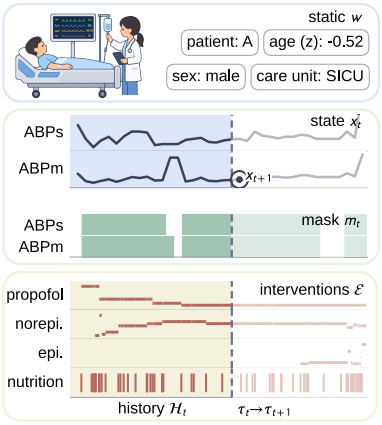


Figure 1: A real ICU stay illustrating the task.

$$\hat{x}_{t+1} = f\big(\mathcal{H}_t,\ \mathcal{E} \cap (\tau_t, \tau_{t+1}],\ \tau_{t+1} - \tau_t\big). \tag{1}$$

Multi-step prediction is a free-running rollout in which predictions are fed back as states while the future intervention course is held to a specified plan, and because that course is an explicit argument of f, separate from the history, substituting an alternative course poses a "what-if" query. The gap between consecutive states, $\tau_{t+1} - \tau_t$, is also an input, since the model must predict across whatever horizon the chart imposes.

Out-of-distribution evaluation. The quantity this paper reports is not the error of f on stays like the ones it was fitted on. A partition cuts the cohort into slices $\mathcal{D}_1, \ldots, \mathcal{D}_K$ along one axis, such as the intensity of a drug class or the age of the patient; a model fitted on $\mathcal{D}_i$ is then scored on the held-out stays of every $\mathcal{D}_j$. The diagonal $i = j$ is the in-distribution (ID) error, the off-diagonal entries $i \neq j$ are the out-of-distribution error, and it is the latter that every result below refers to.

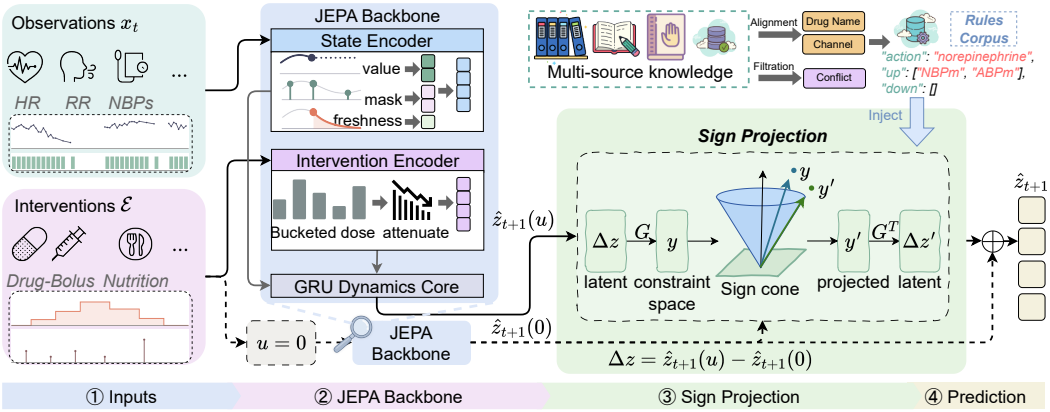


Figure 2: PHYSIOWORLD's forward pass in four stages. (1) States x_t and interventions $\mathcal{E}$ are the inputs. (2) The state and intervention encoders feed a GRU core that runs once under the recorded doses u and once with $u = 0$, giving $\hat{z}_{t+1}(u)$ and $\hat{z}_{t+1}(0)$. (3) Their difference Δz is mapped by G into channel units y and projected onto the sign cone to give $\Pi(y)$. (4) The correction $G^\top(\Pi(y) - y)$ is added back into the latent to give $\hat{z}_{t+1}$.

Stream representation. The intervention stream is a sparse, asynchronous event set, while the network needs a value at every prediction time, so we bridge the two with an *exposure field* $u(s)$, whose entry $u_k(s)$ states how strongly intervention channel k is acting on the patient at continuous time s. A first-order elimination kernel, $u_k(s) = \sum_i v_i e^{-(s-s_i)/\theta_k}$, is the usual choice (Hall & Hall, 2021), but it writes a standard pharmacological model into a representation that this paper otherwise keeps free of pharmacological assumptions, so that any directional structure the model later exhibits can be attributed to the injected rule corpus rather than to this encoding. We therefore instantiate u as a bucketed dose history, summing the dose of each channel within five fixed elapsed-time windows whose boundaries encode a coarse timescale, but the same one for every channel, so that no drug receives a fitted constant. Infusions, boluses and ventilator set-points enter this representation differently, and the state encoder additionally receives a per-channel freshness feature computed from the time since that channel was last measured; Appendix C gives the window boundaries, the treatment of each intervention type, and the freshness definition.

4 PHYSIOWORLD

We now introduce PHYSIOWORLD, which separates what a forward pass predicts from what pharmacology is able to constrain. The separation is what buys the out-of-distribution behavior: the rules the operator enforces are fixed before any record is read, so they are the same under every dose regime, while everything estimated from data is left where the shift can reach it. Three objects are fixed before training begins: the rule corpus, the sign cone it defines at each step, and the projection onto that cone, which has no parameters. Two are learned: the backbone that produces the intervention effect, and the calibration map G in whose rows that effect is read, whose directions an anchor term pins to the channels so that the signs cannot move with the dose regime.

Rule corpus. We inject a corpus of *intervention rules* that state, for each of 28 drugs, the sign/direction of the drug's effect on each channel it touches, 70 (intervention, channel) pairs in all. We keep the sign and omit the magnitude, because dose is titrated against the quantity it changes: a vasopressor is increased as pressure falls and withdrawn as it recovers, so the observed association between dose and effect runs opposite to the drug's action, and the magnitude of the effect cannot be reliably recovered from observational data (Walker, 1996). The direction, by contrast, is a statement about pharmacological mechanism and is unaffected by this confounding variable. Constraints on the state itself, such as per-channel physiological ranges and the ordering systolic $\geq$ mean $\geq$ diastolic, bound which values are possible, but we do not use them, because the failure we target lies in how a prediction *reacts* to an intervention, and that reaction can be wrong while every value stays inside the feasible set. Appendix B lists the corpus and its coverage by channel.

A second forward pass. A rule about a drug constrains what that drug does, but a standard forward pass does not isolate that quantity: it reports the next latent with the intervention's contribution and the patient's own course already summed. We recover that contribution by holding the weights and the recorded observations fixed and advancing the same predictor twice along the stay (Figure 2), once under the recorded dose entries u and once with the dose entry of every covered drug zeroed at every step,

$$\Delta z_{t+1} \;=\; \hat{z}_{t+1}(u) \;-\; \hat{z}_{t+1}(0). \tag{2}$$

Both passes read the same states, so the encodings $z_{1:t}$ are identical and none of the difference originates in the observed course; only the dose entries of the intervention embedding a_s, $s \le t$, of Section 4 differ, and the non-dose intervention channels are untouched. The recurrent states separate at the first administered dose and stay apart, since the intervention encoding enters the recurrence, so Δz_{t+1} is the contribution of the dose history as a whole rather than of the dose delivered in $(\tau_t, \tau_{t+1}]$ alone. That contribution is dominated by the recent past, a trailing window of $w = 3$ steps recovering it to within 1% (Appendix D.2), which is what licenses the cone of Section 4 to be built from the drugs active at t. The two passes advance as parallel recurrent states in one sweep, at the cost of one extra forward pass. Appendix D discusses withdrawing the non-dose channels as well, and severing the intervention stream from the recurrent core so that $u = 0$ would return the untreated course exactly.

The sign projection. Let $\mathcal{K}_t$ be the corpus-covered (drug, channel) pairs whose drug is active at step t. Where every pair in $\mathcal{K}_t$ touching a channel c agrees on the sign of its effect, write σ_c for that common sign; where they disagree, as when a vasopressor and a sedative run together, no sign is implied and none is imposed. In practice, conflict is infrequent: pooled over all channels, it affects 2.3% of scored entries and 4.1% of those under treatment (Appendix B, Table 6). Write $y = G\,\Delta z_{t+1}$ for the effect in channel units. G is a $C \times d$ calibration map with orthonormal rows ($C \le d$) carrying the d-dimensional latents of Section 4 into the C channels, trained by an anchor term alone and separate from the readout head: $\mathcal{L}_{\text{anchor}} = |\mathcal{M}|^{-1} \sum_{(t,c)\in\mathcal{M}} ((Gz_t)_c - \tilde{x}_{t,c})^2$, over the charted entries $\mathcal{M}$ only, with $\tilde{x}$ the standardized channel values. Row c is thus fit to read channel c, which fixes its sign rather than leaving it free: the flipped row predicts $-\tilde{x}_c$ and the term rejects it. Under the default training, no scored row of G reverses across seeds and initializations (Appendix D.1). Because G is linear, an anchor fit on levels transfers to differences without further assumptions, $G(z_1 - z_2) = Gz_1 - Gz_2$, so $(G\Delta z)_c$ reads the same channel that $(Gz)_c$ does. The operator then projects onto the resulting cone and returns the correction to the latent,

$$\Pi(y)_c \;=\; \begin{cases} \sigma_c \max(\sigma_c y_c,\, 0), & \sigma_c \text{ defined}, \\ y_c, & \text{otherwise}, \end{cases} \qquad \hat{z}_{t+1} \;=\; \hat{z}_{t+1}(0) + \Delta z_{t+1} + G^\top(\Pi(y) - y)\,. \tag{3}$$

Write $\Delta z'_{t+1} = \Delta z_{t+1} + G^\top(\Pi(y) - y)$ for the corrected effect, so that $\hat{z}_{t+1} = \hat{z}_{t+1}(0) + \Delta z'_{t+1}$. The corrected $\hat{z}_{t+1}$ is the state advanced into step $t{+}1$, so the correction enters the recurrence rather than acting only at the readout; this is what separates the operator from filtering the decoded output. Algorithm 1 states the per-step procedure in full.

Proposition 1 (Properties of the sign projection)**.** *Let $\mathcal{S} = \{y \in \mathbb{R}^C : \sigma_c y_c \ge 0$ for every c on which σ_c is defined$\}$, and let Π be as in equation 3. Then $\mathcal{S}$ is a nonempty closed convex cone, Π is the Euclidean projection onto it, and for every $y \in \mathbb{R}^C$:*

(i) $\Pi(y) \in \mathcal{S}$;
(ii) $\Pi(y) = y$ *whenever* $y \in \mathcal{S}$;
(iii) $\|\Pi(y) - \Pi(y')\| \le \|y - y'\|$ *for every* y';
(iv) for every rule-consistent $y^\star \in \mathcal{S}$,

$$\|y - y^\star\|^2 - \|\Pi(y) - y^\star\|^2 \;\ge\; \|y - \Pi(y)\|^2.$$

All four properties are elementary, and the proof is in Appendix J.

Remark 1 (Scale)**.** *$\mathcal{S}$ is coordinatewise, so rescaling row c of G by any $\rho_c > 0$ leaves unchanged which components Π clips and no per-cohort calibration of scale is needed. Orthonormal rows fix the size of the write-back as well as its direction, and $GG^\top = I_C$ makes $G\,\Delta z'_{t+1} = \Pi(y)$ exact, so the corrected latent lands in the cone rather than near it.*

Since G trains with the projection active, a G on which the constraint never binds would reduce the operator to the identity at no cost in accuracy; the projection therefore reads a detached G, which

removes rotating G as a descent direction. The remaining route, collapsing Δz, is measured rather than forbidden: the projection moves the response on 54 to 61% of the scored entries on which a sign is defined, and Table 8 varies the initialization, the anchor weight, the orthonormality of the rows and the fitting protocol. The guarantee is also limited: it fixes signs but not magnitudes, it imposes nothing where active drugs disagree or where the corpus is silent, and it holds for $G\,\Delta z$ and not for the decoded prediction, which is why Section 5.4 measures agreement on the decoded output. Convexity is a further reason the rule corpus is stated as signs, since the admissible set of a conditional rule is a union of half-spaces, on which the nearest point need not be unique and non-expansiveness is lost.

JEPA backbone and training. The backbone is a joint-embedding predictive architecture over the represented streams, in which an encoder maps each step's state $[x_t, m_t, \phi_t]$ to a latent z_t and a recurrent core consumes $[z_t, a_t, \delta_t]$. Here a_t embeds the bucketed intervention history introduced in the stream representation, and δ_t contains the gap behind step t and the gap ahead to the target, each on a log and a bounded scale, together with a continuous-time encoding of τ_t, and a predictor then maps the recurrent latent and a_t to $\hat{z}_{t+1}$. Training follows LeJEPA (Balestriero & LeCun, 2025): the target is the same online encoder's encoding of the true next state, with no EMA teacher and no stop-gradient, and collapse is prevented by a distributional regularizer that pushes the embedding distribution toward an isotropic Gaussian, giving $\mathcal{L}_1 = (1-\lambda)\,\|\hat{z}_{t+1} - z_{t+1}\|^2 + \lambda\,\mathcal{L}_{\mathrm{SIGReg}}$. Training has two stages: the first trains the encoder, predictor and intervention pathway without any physical-space loss, and the second freezes the backbone and trains only a readout, so that no physical-space gradient reaches the representation and decoded accuracy can be attributed to what the representation encodes (Bardes et al., 2024). The full objective adds the anchor term of Section 4, $\mathcal{L} = \mathcal{L}_1 + \lambda_{\mathrm{anchor}}\mathcal{L}_{\mathrm{anchor}}$ with $\lambda_{\mathrm{anchor}} = 1$, and carries it through both stages: G is the map the cone is defined on, so it is anchored while the representation forms and again while the readout trains. Inside the operator G enters under stop-gradient, so $\mathcal{L}_1$ trains the backbone through the corrected latent but cannot rotate G to make the projection vacuous; G is updated by $\mathcal{L}_{\mathrm{anchor}}$ alone. The readout decodes a delta on the carry-forward anchor, $\hat{x}_{t+1} = x_t + \mathrm{MLP}(\hat{z}_{t+1})$, which is the same output parameterization every baseline uses.

5 Experiments

Section 5.2 asks which kind of distribution shift damages these models and whether the pharmacology repairs it, Section 5.3 whether the damage compounds when a model runs on its own predictions, Section 5.4 whether the dose is used in the direction pharmacology states, and Section 5.5 where the gain comes from, by varying the injected corpus and by removing one component at a time.

5.1 Setup

Datasets. We evaluate five cohorts from three databases, selected on the primary discharge diagnosis so that every stay belongs to exactly one. **MIMIC-IV** (Johnson et al., 2023) covers 2008 to 2019 under MetaVision charting and contributes circulatory (20,128 stays) and injury (7,746) cohorts; **MIMIC-III** (Johnson et al., 2016), restricted to its CareVue era of 2001 to 2008, contributes 10,020 and 4,480 stays for the same two chapters and is independent of MIMIC-IV in charting system, period and patients; **VitalDB** (Lee et al., 2022) adds 5,391 surgical cases from a tertiary center in Seoul, gridded at five minutes rather than one hour. The same corpus is applied to all five without modification, which tests whether a stated direction can be *reused as written*.

Task. The task is the one-step prediction of Section 3 for the nine vital-sign channels defined in Appendix B. The intervention stream covers drugs, blood products and ventilator set-points, with 35 channels on MIMIC-IV, 30 on MIMIC-III and 12 on VitalDB, and predictions are scored only on freshly measured channels whose carry-forward anchor was itself measured.

Shift axes. Axes are built by a *split-then-slice* protocol: following Gulrajani & Lopez-Paz (2020), the cohort is partitioned once by a global 80/20 subject-wise split, each slice is the intersection of a stratum with that split, and training sets are downsampled to the smallest slice so that data volume does not confound the comparison. *Population axes* stratify patients by characteristics fixed on admission; *intervention-intensity axes* band exposed stays by within-class dose rank, taking the bottom and top quartiles of burden, which assumes no equipotency across drugs. There are five:

Table 1: In-distribution and out-of-distribution RMSE (↓) on MIMIC-IV, physical units, mean ± std over five seeds. The ID column is a model trained on the full training split and tested on the full test split. Each OOD entry averages the off-diagonal of that axis's train-slice × test-slice matrix. **Bold**: best per shift axis. Underline: best external baseline. The ID column is unmarked by design.

		Population shift			Intervention-intensity shift	
Method	In-dist	Age	Disease	Era	Vasopressor	Sedative
Reference and data-driven forecasters						
Persistence	10.57	10.86	11.07	10.86	10.83	10.91
PatchTST	8.93 ± 0.05	9.68 ± 0.07	9.18 ± 0.07	8.92 ± 0.06	12.15 ± 0.08	11.60 ± 0.08
GraFITi	8.95 ± 0.05	9.62 ± 0.07	9.21 ± 0.07	8.95 ± 0.06	11.85 ± 0.08	11.42 ± 0.07
Chronos-2	8.88 ± 0.05	9.55 ± 0.06	9.12 ± 0.06	8.86 ± 0.06	11.70 ± 0.08	11.35 ± 0.07
CRN	9.34 ± 0.06	9.85 ± 0.07	9.45 ± 0.07	9.20 ± 0.06	10.95 ± 0.08	10.68 ± 0.07
Causal Transformer	9.08 ± 0.05	9.72 ± 0.06	9.30 ± 0.06	9.05 ± 0.06	10.45 ± 0.07	10.28 ± 0.07
Robustness by reweighting or invariance						
IPTW	8.12 ± 0.06	9.65 ± 0.07	10.55 ± 0.08	9.40 ± 0.07	10.55 ± 0.07	10.38 ± 0.07
IRM	9.31 ± 0.06	10.75 ± 0.09	11.12 ± 0.09	10.65 ± 0.08	12.35 ± 0.10	12.10 ± 0.09
CORAL	10.21 ± 0.07	10.35 ± 0.08	10.48 ± 0.08	10.15 ± 0.07	11.20 ± 0.08	10.95 ± 0.08
World models						
JEPA	7.95 ± 0.04	9.34 ± 0.08	10.68 ± 0.08	9.46 ± 0.07	11.15 ± 0.08	11.60 ± 0.08
Soft constraint	7.90 ± 0.04	9.25 ± 0.06	8.85 ± 0.06	8.62 ± 0.05	11.10 ± 0.07	10.75 ± 0.07
Clamp	7.92 ± 0.04	8.55 ± 0.05	8.40 ± 0.05	8.20 ± 0.04	9.65 ± 0.06	9.38 ± 0.06
PHYSIOWORLD	7.95 ± 0.04	$\mathbf{8.41 \pm 0.05}$	$\mathbf{8.26 \pm 0.05}$	$\mathbf{8.15 \pm 0.04}$	$\mathbf{9.42 \pm 0.06}$	$\mathbf{9.21 \pm 0.06}$

age, disease subtype and era, cut into 5, 4 and 4 slices, and vasopressor and sedative. VitalDB has too few exposed stays for the vasopressor axis, so cross-database comparison uses the sedative axis alone (Appendix G). Test slices differ in intrinsic difficulty (Koh et al., 2021), so carry-forward persistence, which scores zero skill by definition, is the reference for the ID column.

Baselines. All methods share the same grid and carry-forward anchor (Zeng et al., 2023; Qiu et al., 2024). The forecasters are PatchTST (Nie et al., 2022), GraFITi (Yalavarthi et al., 2024) for irregular sampling, and the pretrained Chronos-2 (Ansari et al., 2025), all three receiving the planned intervention course as a known-future covariate. CRN (Bica et al., 2020) and the Causal Transformer (Melnychuk et al., 2022) forecast under counterfactual time-varying interventions, the setting closest to ours, while IPTW reweights training stays by inverse treatment probability and IRM (Arjovsky et al., 2019) and CORAL (Sun & Saenko, 2016) seek features invariant across training slices. JEPA, Soft constraint and Clamp share our backbone and differ only in where the constraint acts: *nowhere, in the loss, or on the output*.

Metrics. We report RMSE in physical units, in and out of distribution, as an unweighted macro mean over channels (Zhang et al., 2024). A *skill score* against persistence, $1 - \mathrm{RMSE}/\mathrm{RMSE}_{\mathrm{pers}}$, is computed within each slice pair and aggregated geometrically (Hamill & Juras, 2006; Davydenko & Fildes, 2013); it is the primary metric across databases, whose sampling rates make absolute RMSE incomparable. We also report *dispersion*, the ratio of the across-patient standard deviation of a model's predictions to that of the truth, where values below one indicate collapse toward the population mean. All numbers are mean and standard deviation over five seeds. Differences against Clamp, the strongest method sharing our backbone, are tested with a patient-level paired bootstrap and a one-sided Wilcoxon signed-rank test over train/test slice pairs (Appendix K).

5.2 SHIFT AXES: POPULATION AGAINST INTERVENTION INTENSITY

Population shift. Across *Age*, *Disease* and *Era*, most models degrade moderately, which is consistent with demographic shift acting mainly on the covariate distribution $P(X)$ while the physiological transition dynamics stay largely invariant. Model capacity appears sufficient here, and the margin of PHYSIOWORLD over Clamp, the strongest baseline, is *consistent but modest*.

Intervention-intensity shift. Under *Vasopressor* and *Sedative* shift the forecasters separate into two groups: **pure sequence models fall *below* persistence** (PatchTST 12.15, persistence 10.83), while methods that model the intervention explicitly remain competitive with it (Causal Transformer, 10.45). As explained in the Introduction, clinical dosing is confounded by indication, so a model fit to observational association reads dose as a marker of severity, and that the two groups diverge under the same shift suggests that *where* the intervention enters the model matters more than model capacity. Domain-invariant objectives do not close the gap and fail in opposite ways (Section 5.3).

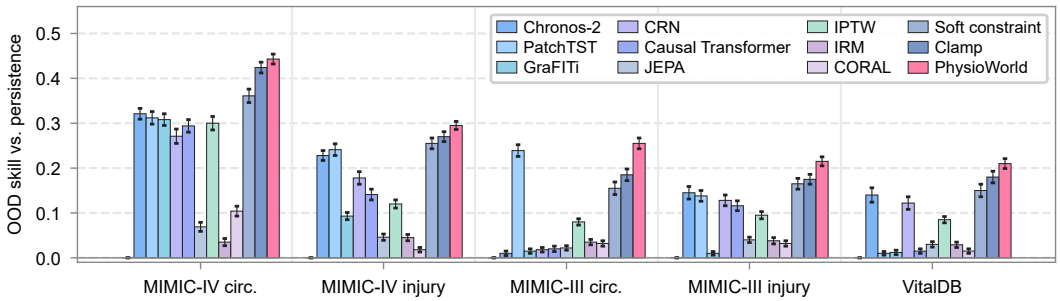


Figure 3: Cross-database generalization under sedative intervention-intensity shift. Skill score is the fractional RMSE reduction over carry-forward persistence; higher is better, 0 means no gain.

Method	Slope ↓	Disp. (→ 1)
Persistence	+0.084	0.985
PatchTST	+0.127	0.631
GraFITi	+0.116	0.700
Chronos-2	+0.124	0.645
CRN	+0.108	0.758
Causal Transformer	+0.115	0.663
IPTW	+0.296	1.825
IRM	+0.472	1.757
CORAL	+0.102	0.415
JEPA	+0.150	0.748
Soft constraint	+0.141	0.765
Clamp	+0.116	0.792
PHYSIOWORLD	**+0.090**	**0.842**

Table 2: Rollout endpoints on MIMIC-IV.

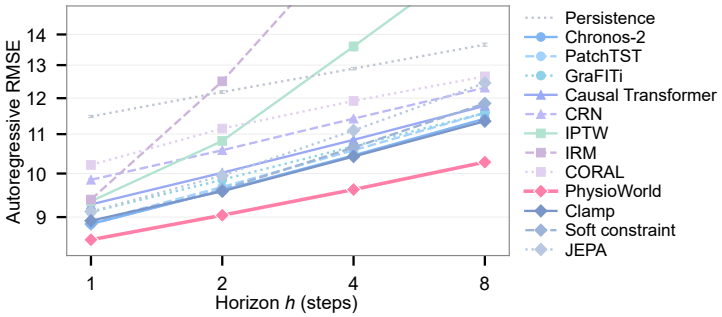


Figure 4: Free-running autoregressive rollout RMSE over an eight-step horizon on MIMIC-IV.

How the knowledge is injected. A soft monotonicity penalty appears to be dominated by the reconstruction gradients and does not bind out of distribution (11.10), and clamping is stronger (9.65) but acts *only on the emitted prediction*, leaving unchanged the latent representation that produced the violation. PHYSIOWORLD, which constrains the signed dose–effect pathway *at every recurrent transition*, is best on all five axes, with in-distribution error equal to its backbone's and a smallest margin of 0.05 over Clamp on *Era*. Figure 3 extends the comparison to five cohorts under sedative shift, where the gain is small but present on every cohort: every data-driven forecaster drops to within 0.02 of persistence on at least one cohort, and which cohort it is differs by method. PHYSIOWORLD stays at or above 0.21 on all five cohorts and exceeds the best baseline on each; its margin over Clamp is positive everywhere but within about two seed standard deviations on four of the five, and the bootstrap interval of the difference excludes zero on four of the five axes (Appendix K). We therefore claim *a consistent ordering, not a win on any individual cohort*.

5.3 HORIZON ANALYSIS UNDER FREE-RUNNING ROLLOUT

After a half-stay warmup, each model's predictions are fed back as its own inputs for eight steps, while the recorded doses continue to be supplied. Because cohort and split are held fixed, the only departure from the training distribution is *the model's own error*, which becomes its input and accumulates with the horizon, so single-step errors are not comparable to the ID column of Table 1. Figure 4 and Table 2 show the error and its growth, with every horizon in Appendix L. PHYSIOWORLD has the lowest error at every horizon, from 8.52 at h=1 to 10.28 at h=8, and the flattest growth among learned models, with a slope of +0.090 against +0.127 for PatchTST. Chronos-2 and PatchTST are the strongest external baselines at h=1 but *compound fastest* among the forecasters and lose roughly a third of their skill over persistence across the rollout, and clamping loses a quarter, whereas PHYSIOWORLD keeps nearly all of its skill (0.258 at h=1, 0.247 at h=8). CORAL's flat slope (+0.102) reflects collapse toward the population mean (dispersion 0.415) rather than robustness, while IRM and IPTW fail in the opposite direction (1.757 and 1.825) and are the only methods whose error diverges. PHYSIOWORLD, by contrast, has the dispersion closest to one of any learned model (0.842), indicating its low error is *not obtained by purely predicting the average patient*.

5.4 DOSE RESPONSE UNDER PERTURBATION

Low error alone cannot separate a model that reads the dose as a drug from one that reads it as a marker of severity. We therefore perturb the doses at test time on the pooled test slices of the two intervention-intensity axes and measure how the one-step prediction $\hat{x}_{t+1}$ responds (Table 3). Two perturbations are applied separately: the doses over the predicted interval $(\tau_t, \tau_{t+1}]$ are set to zero and the dose courses are permuted across patients within a test slice, which keeps their distribution and breaks their link to the patient. *Sensitivity* is the mean absolute change in $\hat{x}_{t+1}$ over covered channels and asks whether the dose is used at all. *Agreement* is the share of covered channel-steps on which that change has the sign $\sigma_c \cdot \mathrm{sign}(\Delta u)$, counted only where the changed doses imply a single sign as in Section 4, and asks whether it is used in the right direction. Agreement is measured on the decoded $\hat{x}$, where the operator gives no guarantee, so it is *not satisfied by construction*. PHYSIOWORLD agrees on 86% of eligible channel-steps under zeroing and 79% under permutation, against at most 77% and 70% respectively for any baseline, while its sensitivity is comparable to that of the forecasters, so the agreement is *not obtained by responding less*. The sequence forecasters agree on fewer than half of eligible channel-steps: given a higher dose they *move the channel against the drug's action*, which is the confounding by indication of Section 4 *observed in the models rather than assumed*.

	Zeroed		Permuted	
Method	Sens.	Agr. %	Sens.	Agr. %
Persistence	0	n/a	0	n/a
PatchTST	1.10	36	0.66	41
GraFITi	1.02	39	0.61	43
Chronos-2	1.15	41	0.70	44
CRN	0.95	58	0.57	55
Causal Transformer	1.05	63	0.63	59
IPTW	0.98	61	0.59	57
IRM	1.60	47	0.95	48
CORAL	0.22	51	0.13	51
JEPA	1.08	46	0.65	47
Soft constraint	1.00	55	0.60	53
Clamp	0.57	77	0.34	70
PHYSIOWORLD	0.93	**86**	0.55	**79**

Table 3: Dose perturbation on the intervention-intensity test slices of MIMIC-IV circulatory.

5.5 KNOWLEDGE AND COMPONENT ABLATIONS

Skill decreases monotonically as more of the stated directions are inverted (Table 4), from 0.198 at full agreement to -0.058 when every direction is reversed, below the unconstrained backbone at 0.053, so **a wrong corpus is worse than none**. Because the operator is the same in every arm, the gain must come from *the content of the knowledge*, and withholding pairs degrades skill more gradually, to 0.152 at half the corpus and 0.107 at a quarter. The two corpora estimated from the record both lose most of the gain: a magnitude band, which asserts a size that no source supplies, falls to 0.093 despite agreeing with pharmacology on every direction, and mined directions agree with pharmacology on only 48% of the 70 pairs, *no better than chance*, and return 0.005. Consequently, the directions this construction needs cannot be recovered from a record in which dose is titrated against the quantity it changes. Permuting the drug labels changes only which agent each direction names and returns 0.066 against the backbone's 0.053 while covering 28.3% of scored entries rather than 13.0%, so *neither the constraint itself nor its reach produces the gain*. Removing the sign projection or the second forward pass lowers the skill of the full method from +0.20 to +0.11 and +0.08 respectively, against +0.05 for the backbone, so the gain comes from *their combination* (Appendix E).

Corpus	Pairs	Cover.	OOD skill
No rules (JEPA)	0	0%	0.053
How much is stated			
PHYSIOWORLD	70	13.0%	0.198
35, stratified	35	7.9%	0.152
17, stratified	17	4.1%	0.107
How much of it is right			
25% inverted	70	13.0%	0.121
50% inverted	70	13.0%	0.024
All inverted	70	13.0%	-0.058
Which drug does what			
Drug labels permuted	70	28.3%	0.066
Where it comes from			
Mined directions	70	13.0%	0.005
What it asserts			
Signs + magnitude band	70	13.0%	0.093

Table 4: Rule corpora, all injected through the same operator. *Cover.*: share of scored entries with a defined sign.

6 CONCLUSION

A clinical world model is used where treatment practice differs from its training records, and its error there is what this paper set out to reduce. PHYSIOWORLD reduces it by isolating an intervention's effect with a second forward pass and constraining its sign to the stated pharmacological direction. The gain is largest under shifts in treatment intensity and holds across five axes and three databases; an inverted or data-mined corpus performs worse than no corpus, so what generalizes is the pharmacology and not the constraint.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this paper, including drafting and revising prose in the method and appendix sections. They were not used to generate or label data, to run experiments, to select which results to report, or to decide the directions in the rule corpus, which are taken from receptor pharmacology and were fixed before any model was trained. Every reported number is the output of a recorded run on real data. The authors verified all content and take full responsibility for the manuscript and the released artifacts.

ETHICS STATEMENT

This work uses existing de-identified critical-care databases and collects no new data from human subjects. MIMIC-IV and MIMIC-III require credentialed access through PhysioNet under a data use agreement, so the released repository contains the cohort-extraction code rather than the data; VitalDB is openly available. A processed cohort is still patient data and is excluded from the release. The rule corpus states directions of drug action and is not clinical guidance. A model that satisfies the constraint has not thereby been validated for any decision about a patient: the operator fixes the sign of a modelled response and says nothing about its magnitude, its timing or its safety, and the guarantee is a statement about the model's own account of a treatment rather than about what the treatment would do, since the counterfactual arm is never observed. The constraint also enforces whatever signs it is given, so a direction stated wrongly would be imposed as firmly as a correct one; the corpus should be reviewed by clinicians before the method is applied in a new setting, and Appendix B reports how much of the scored data it reaches. Any use on patient data would require prospective evaluation in the intended clinical setting.

REPRODUCIBILITY STATEMENT

Section 5.1 specifies the cohorts, the five shift axes, the baselines and the metrics; Section 4 specifies the model, the training procedure and the operator calibration. Appendix F lists all hyperparameters and Appendix G the holdout construction. MIMIC-IV is available to any credentialed researcher through PhysioNet, and the code, the rule corpus as a machine-readable file in both JSON and CSV, and the hyperparameter and holdout documents are available for review at https://anonymous.4open.science/r/physioworld-A2B7/.

REFERENCES

Denis Agniel, Isaac S Kohane, and Griffin M Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 361, 2018. ISSN 0959-8138. doi: 10.1136/bmj.k1479. URL https://www.bmj.com/content/361/bmj.k1479.

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International conference on machine learning*, pp. 136–145. PMLR, 2017.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Anand Avati, Martin Seneviratne, Emily Xue, Zhen Xu, Balaji Lakshminarayanan, and Andrew M Dai. Beds-bench: Behavior of ehr-models under distributional shift–a benchmark. *arXiv preprint arXiv:2107.08189*, 2021.

Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022.

Marco Bagatella, Matteo Pirotta, Ahmed Touati, Alessandro Lazaric, and Andrea Tirinzoni. Td-jepa: Latent-predictive representations for zero-shot reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 36211–36244, 2026.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela Van Der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv preprint arXiv:2002.04083*, 2020.

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

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. In *International Conference on Learning Representations*, volume 2026, pp. 89791–89807, 2026.

Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020.

Gal Dalal, Krishnamurthy Dvijotham, Matej Vecerik, Todd Hester, Cosmin Paduraru, and Yuval Tassa. Safe exploration in continuous action spaces. *arXiv preprint arXiv:1801.08757*, 2018.

Andrey Davydenko and Robert Fildes. Measuring forecasting accuracy: The case of judgmental adjustments to sku-level demand forecasts. *International Journal of Forecasting*, 29(3):510–522, 2013.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. *Advances in neural information processing systems*, 32, 2019.

Priya L Donti, David Rolnick, and J Zico Kolter. Dc3: A learning method for optimization with hard constraints. *arXiv preprint arXiv:2104.12225*, 2021.

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

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in neural information processing systems*, 32, 2019.

Xiao Gu, Yu Liu, Zaineb Mohsin, Jonathan Bedford, Anshul Thakur, Peter J Watkinson, Lei A Clifton, Tingting Zhu, and David A Clifton. Are time series foundation models ready for vital sign forecasting in healthcare? In *ML4H@ NeurIPS*, pp. 401–419, 2024.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

John E. Hall and Michael E. Hall. *Guyton and Hall Textbook of Medical Physiology*. Elsevier, Philadelphia, PA, 14th edition, 2021. ISBN 978-0-323-59712-8.

Thomas M Hamill and Josip Juras. Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 132(621C):2905–2923, 2006.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Miguel A Hernán and James M Robins. *Causal inference*. CRC Boca Raton, FL, 2010.

Nils Hinrichs, Tobias Roeschl, Pia Lanmueller, Felix Balzer, Carsten Eickhoff, Benjamin O'Brien, Volkmar Falk, and Alexander Meyer. Short-term vital parameter forecasting in the intensive care unit: A benchmark study leveraging data from patients after cardiothoracic surgery. *PLOS Digital Health*, 3(9):e0000598, 2024.

Yongchao Huang and Hassan Raza. Knowledge, rules and their embeddings: Two paths towards neuro-symbolic jepa. *arXiv preprint arXiv:2603.13265*, 2026.

Zeshan M Hussain, Rahul G Krishnan, and David Sontag. Neural pharmacodynamic state space modeling. In *International Conference on Machine Learning*, pp. 4500–4510. PMLR, 2021.

Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3):364–373, 2020. ISSN 1546-170X. doi: 10.1038/s41591-020-0789-4. URL https://doi.org/10.1038/s41591-020-0789-4.

John PA Ioannidis, Anna-Bettina Haidich, Maroudia Pappa, Nikos Pantazis, Styliani I Kokori, Maria G Tektonidou, Despina G Contopoulos-Ioannidis, and Joseph Lau. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *Jama*, 286(7):821–830, 2001.

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):1, 2023. doi: 10.1038/s41597-022-01899-x.

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

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0213-5. URL https://doi.org/10.1038/s41591-018-0213-5.

Yuki Kotani, Annamaria Di Gioia, Giovanni Landoni, Alessandro Belletti, and Ashish K Khanna. An updated "norepinephrine equivalent" score in intensive care as a marker of shock severity. *Critical Care*, 27(1):29, 2023.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, 2022.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, et al. Mira: Medical time series foundation model for real-world health data. *Advances in neural information processing systems*, 38:98657–98685, 2026.

Ke Liu, Mengxuan Li, Yanyi Bao, Tianyun Zhang, Chong Chu, Jiajun Bu, and Haishuai Wang. Medical world models: representing medical states, modelling clinical dynamics and guiding intervention policies. *arXiv preprint arXiv:2606.16721*, 2026a.

Weixin Liu, Juming Xiong, Congning Ni, Yanfan Zhu, Xingtao Lin, Bradley A Malin, and Zhijun Yin. Drift: Direct-recursive intervention-conditioned forecasting of icu physiological trajectories. *arXiv preprint arXiv:2607.25864*, 2026b.

Xingchao Liu, Xing Han, Na Zhang, and Qiang Liu. Certified monotonic neural networks. *Advances in Neural Information Processing Systems*, 33:15427–15438, 2020.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International conference on learning representations*, volume 2024, pp. 11116–11140, 2024.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International conference on machine learning*, pp. 15293–15329. PMLR, 2022.

Marcel Meyer, Sascha Kaltenpoth, Kevin Zalipski, and Oliver Müller. Rethinking evaluation in the era of time series foundation models:(un) known information leakage challenges. *arXiv preprint arXiv:2510.13654*, 2025.

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. *arXiv preprint arXiv:2602.03569*, 2026.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Unnseo Park, Venkatesh Sivaraman, and Adam Perer. How consistent are clinicians? evaluating the predictability of sepsis disease progression with dynamics models. *arXiv preprint arXiv:2404.07148*, 2024.

Gernot Pucher, Amin Dada, Aurel Agbodoyetin, Felix Nensa, Martin Schuler, H Christian Reinhardt, Jens Kleesiek, and Christopher M Sauer. Can one model fit all? evaluating foundation models for time series forecasting across clinical medicine. *Artificial Intelligence in Medicine*, pp. 103473, 2026.

Zhaozhi Qian, William Zame, Lucas Fleuren, Paul Elbers, and Mihaela van der Schaar. Integrating expert odes into neural odes: Pharmacology and disease progression. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11364–11383. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5ea1649a31336092c05438df996a3e59-Paper.pdf.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S Jensen, Zhenli Sheng, et al. Tfb: Towards comprehensive and fair benchmarking of time series forecasting methods. *arXiv preprint arXiv:2403.20150*, 2024.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. *Advances in neural information processing systems*, 32, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Joseph Sill. Monotonic networks. *Advances in neural information processing systems*, 10, 1997.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Alexander M Walker. Confounding by indication. *Epidemiology*, 7(4):335–336, 1996.

Jiangyuan Wang, Xuyong Chen, Junwei He, Xu Xu, Shasha Xie, and Fuman Han. Chronomedical-world: A medical world model for learning patient trajectories from longitudinal care data. *arXiv preprint arXiv:2605.21963*, 2026.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in neural information processing systems*, 37:469–498, 2024.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A semantic loss function for deep learning with symbolic knowledge. In *International conference on machine learning*, pp. 5502–5511. PMLR, 2018.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. Grafiti: Graphs for forecasting irregularly sampled time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16255–16263, 2024.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, et al. Medical world model: Generative simulation of tumor evolution for treatment planning. *arXiv preprint arXiv:2506.02327*, 2025.

Yixuan Yang, Mehak Arora, Ryan Zhang, Baraa Abed, Junseob Kim, Tilendra Choudhary, Md Hassanuzzaman, Kevin Zhu, Ayman Ali, Chengkun Yang, et al. Clin-jepa: A multi-phase co-training framework for joint-embedding predictive pretraining on ehr patient trajectories. *arXiv preprint arXiv:2605.10840*, 2026.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*, 2023.

Seungil You, David Ding, Kevin Canini, Jan Pfeifer, and Maya Gupta. Deep lattice networks and partial monotonic functions. *Advances in neural information processing systems*, 30, 2017.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *ICML*, pp. 60179–60196, 2024.

Algorithm 1 One prediction step of PHYSIOWORLD. The two passes are advanced as parallel recurrent states over a single sweep of the stay; they share the encoded state and rejoin only at the correction.

Require: encoder E, recurrent core f, predictor P, calibration map G, rule corpus $\mathcal{R}$; states h^u_{t-1}, h^0_{t-1}

$z_t \leftarrow E([x_t, m_t, \phi_t])$ ▷ observed; shared by both passes
$a^u_t \leftarrow$ intervention embedding under the recorded doses
$a^0_t \leftarrow a^u_t$ with the dose entry of every covered drug set to 0
$h^u_t \leftarrow f(h^u_{t-1}, [z_t, a^u_t, \delta_t])$
$h^0_t \leftarrow f(h^0_{t-1}, [z_t, a^0_t, \delta_t])$
$\hat{z}_{t+1}(u) \leftarrow P(h^u_t, a^u_t)$
$\hat{z}_{t+1}(0) \leftarrow P(h^0_t, a^0_t)$
$\Delta z_{t+1} \leftarrow \hat{z}_{t+1}(u) - \hat{z}_{t+1}(0)$
$y \leftarrow G\,\Delta z_{t+1}$
for all channels c **do**
    $\sigma_c \leftarrow$ the common sign in $\mathcal{R}$ over the drugs active at t that touch c, if they agree; undefined otherwise
    $\Pi(y)_c \leftarrow \sigma_c \max(\sigma_c y_c, 0)$ if σ_c is defined, else y_c
end for
$\hat{z}_{t+1} \leftarrow \hat{z}_{t+1}(0) + \Delta z_{t+1} + G^\top(\Pi(y) - y)$
$\hat{x}_{t+1} \leftarrow x_t + \mathrm{MLP}(\hat{z}_{t+1})$

Next step. $z_{t+1} \leftarrow E([x_{t+1}, m_{t+1}, \phi_{t+1}])$ when the next state is observed, and $z_{t+1} \leftarrow \hat{z}_{t+1}$ under free-running rollout, so the corrected latent, not the decoded output, is what both recurrent states read next.

A THE FORWARD PASS, STEP BY STEP

Algorithm 1 states the per-step computation of Section 4 and Section 4 in full.

B RULE CORPUS

The corpus consists of 70 (intervention, channel, sign) triples over 28 drugs, listed in Table 7, where only signs are stated and magnitudes are excluded for the reason given for the rule corpus in Section 4. Table 5 gives, for each drug, the number of channels it has rules on, the scored entries while it is active, and its coverage, which ranges from 9.5% to 18.0%, so no single drug dominates the constraint. Drugs are grouped by the classes that define the intervention-intensity axes of Section 5.1, so each rule can be matched to the axis on which it is evaluated. Four drugs have rules in both directions, and all four are intended: dobutamine raises systolic pressure and lowers diastolic pressure, and hydralazine, milrinone and phenylephrine each combine their pressure effect with an opposite reflex change in heart rate. The channels are heart rate (HR), respiratory rate (RR), peripheral oxygen saturation (SpO_2), and systolic, diastolic and mean blood pressure, measured non-invasively (NBPs, NBPd, NBPm) or from an arterial line (ABPs, ABPd, ABPm).

Table 6 reports, for each channel c, how often active drugs leave it unconstrained. *Rules* is the number of corpus rules that name c. *Conflict* is the share of all scored entries of c on which active drugs have rules on c that disagree, so that no sign is imposed, and *Silent* is the share on which no active drug has a rule on c, whether because no corpus drug is active or because the active ones do not touch c. *Conflict/treated* divides the conflict share by the share of entries at which at least one corpus drug is active, and so measures how often treatment, when present, leaves the channel unconstrained. The *Pooled* row aggregates over the scored entries of all nine channels.

Coverage per channel. The 70 rules are unevenly spread over the nine channels. Rules per channel, as raising / lowering (derived from Table 7):

HR	RR	SpO_2	NBPs	NBPd	NBPm	ABPs	ABPd	ABPm
5/9	0/8	0/0	1/0	0/1	8/15	1/0	0/1	8/13

Table 5: Per-drug statistics: number of channels the drug has rules on, scored entries while it is active, and rule coverage (Cover.), the share of those entries with a defined sign.

Drug	Channels	Scored entries	Cover.
propofol	3	341,457	14.2%
fluid_bolus	2	236,013	11.5%
phenylephrine	3	207,196	15.0%
fentanyl	2	199,835	12.8%
morphine	3	154,708	13.5%
furosemide	2	100,684	10.0%
midazolam	3	85,647	14.8%
hydralazine	3	82,021	12.0%
metoprolol	3	81,694	13.2%
epinephrine	3	72,454	11.8%
nitroglycerin	2	65,831	12.5%
dexmedetomidine	3	63,616	14.0%
prbc	2	63,132	10.5%
norepinephrine	2	61,495	13.0%
amiodarone	3	53,497	11.2%
hydromorphone	3	50,248	15.2%
labetalol	3	33,263	12.2%
ffp	2	33,101	10.8%
nicardipine	2	32,913	9.5%
milrinone	3	31,606	14.5%
dopamine	3	29,113	12.6%
dobutamine	5	27,704	16.0%
lorazepam	1	25,286	17.0%
vasopressin	2	19,915	11.0%
esmolol	3	13,091	13.5%
cisatracurium	1	10,131	17.5%
nitroprusside	2	5,288	10.2%
diazepam	1	545	18.0%

Table 6: Channel-level rule counts, conflict and silence on MIMIC-IV circulatory.

Channel	Rules	Conflict	Silent	Conflict/treated
RR	8	0.0%	82.5%	0.0%
HR	14	4.4%	76.6%	8.7%
NBPm	23	2.7%	85.2%	5.2%
ABPm	21	5.6%	81.4%	12.4%
NBPs/d, ABPs/d	4	0.0%	99.0%	0.0%
SpO2	0	0.0%	100.0%	0.0%
Pooled	70	2.3%	84.7%	4.1%

Three channels carry 58 of the 70 rules, SpO_2 carries none, and the four systolic and diastolic channels carry one each, which limits the operator in two ways. Where no rule covers a channel, the operator imposes no constraint, and the three densely covered channels are also the ones where a raising and a lowering drug can be active together, in which case no sign is implied and again no constraint is imposed; propofol with norepinephrine is a common combination that produces this case.

C Stream Representation

Aligned state rows are 2.99M and 1.24M for the MIMIC-IV circulatory and injury cohorts, 1.27M and 0.57M for the MIMIC-III cohorts, and 1.00M for VitalDB. The exposure field of Section 3 is a bucketed dose history, in which the dose of each intervention channel is summed within five fixed elapsed-time windows (0–0.5 h, 0.5–2 h, 2–6 h, 6–24 h, and >24 h), giving an $A \times 5$ matrix at every prediction time. Infusions need different handling from boluses, because a charted infusion row restates the current rate and does not add a dose, so each rate channel holds its most recent value

Table 7: The full set of intervention rules: the channels each drug raises and lowers. Only signs are stated; no magnitude is asserted.

Drug	Raises	Lowers	Drug	Raises	Lowers
Vasopressors and inotropes			*Rate control and antiarrhythmics*		
norepinephrine	NBPm, ABPm		metoprolol		HR, NBPm, ABPm
epinephrine	HR, NBPm, ABPm		esmolol		HR, NBPm, ABPm
phenylephrine	NBPm, ABPm	HR	labetalol		HR, NBPm, ABPm
vasopressin	NBPm, ABPm		amiodarone		HR, NBPm, ABPm
dopamine	HR, NBPm, ABPm		*Vasodilators*		
dobutamine	HR, NBPs, ABPs	NBPd, ABPd	nitroglycerin		NBPm, ABPm
milrinone	HR	NBPm, ABPm	nicardipine		NBPm, ABPm
Sedatives and opioids			nitroprusside		NBPm, ABPm
propofol		NBPm, ABPm, RR	hydralazine	HR	NBPm, ABPm
fentanyl		RR, HR	*Volume, blood products and other*		
midazolam		RR, NBPm, ABPm	fluid_bolus	NBPm, ABPm	
dexmedetomidine		HR, NBPm, ABPm	prbc	NBPm, ABPm	
lorazepam		RR	ffp	NBPm, ABPm	
diazepam		RR	furosemide		NBPm, ABPm
morphine		RR, NBPm, HR	cisatracurium		RR
hydromorphone		RR, NBPm, HR			

(zero-order hold) and is integrated over each window to give the administered dose. Treating such rows as impulses would instead make exposure scale with charting frequency, which tracks acuity (Agniel et al., 2018), and the model would learn nursing behavior in place of pharmacology, while ventilator set-points, which carry no dose semantics, bypass the bucketing as their current held value. On the state side, the encoder receives $[x_t, m_t, \phi_t]$, where $\phi_{t,c} = \exp(-\delta_{t,c}/6\text{h})$ is a per-channel freshness computed from the time since channel c was last measured. This is a staleness weight on observed information and involves no pharmacokinetic modeling: its time constant is fixed, and half-weight is reached at $6 \ln 2 \approx 4.2$ h. The first-order elimination kernel that the bucketed history replaces is also a poor prior when its constant θ_k is left free: fitting it end-to-end yields 4.7 h for norepinephrine, whose elimination half-life is about two minutes.

D Design Choices and Limits

D.1 Calibration map sensitivity

Table 8 varies the initialization of G, the weight of the anchor term, the orthonormality of its rows, and the protocol under which it is fitted, over five seeds each. *OOD RMSE* is averaged over the five shift axes of Table 1, as in Table 9, so the Default row is PhysioWorld itself. *Binding rate* is the fraction of the (step, channel) pairs on which a sign is defined that the projection clips, as a range across the five OOD cohorts. *Stability* is the within-run cosine between row c at 80% and at 100% of training, averaged over scored channels, and reports whether G is still moving when training stops. *Seed agreement* is the between-seed cosine of row c, as median and minimum over scored channels, and reports whether the anchor term identifies a single set of directions; two unrelated unit vectors in $\mathbb{R}^d$ agree to $\sqrt{2/\pi d} \approx 0.14$ on average, and the Control row sits at or below that level. *Sign flips* counts scored channels on which two seeds disagree in sign, out of the 9 scored channels. $\|GG^\top - I\|_\infty$ measures departure from orthonormal rows.

Four readings follow. First, the directions are set by the anchor term and not by where optimization starts: rows initialized along the coordinate axes reach the same solution as the default to a median cosine of 0.97, with no channel reversing sign and OOD error unchanged at 8.70. Second, the anchor term is what does that work. Weakening it tenfold costs 0.20 RMSE, drops between-seed agreement to 0.71 with two channels reversing, and pulls the binding rate toward the 50% a sign-uninformative G would produce; strengthening it tenfold changes nothing, so the default sits on a plateau rather than at a tuned point. Third, orthonormal rows matter for the write-back and not for the decision: leaving them unnormalized leaves the binding rate intact, since which components Π clips depends only on signs (Remark 1), but costs 0.25 RMSE, because $GG^\top \neq I$ leaves the corrected latent

Table 8: Sensitivity of the calibration map G to initialization, anchor loss weight, parameterization and fitting protocol; columns are defined in the text and all row statistics are restricted to the 9 scored channels.

Variant	OOD RMSE (↓)	Binding rate	Stability (→ 1)	Seed agreement med / min	Sign flips (↓)	$\|GG^\top - I\|_\infty$
Default	8.69	54–61%	0.999	0.98/0.94	0/9	$< 10^{-6}$
Initialization						
Axis-aligned rows	8.70	53–60%	0.999	0.97/0.92	0/9	$< 10^{-6}$
Anchor loss weight						
$\lambda_{\text{anchor}} \times 0.1$	8.89	50–53%	0.98	0.71/0.38	2/9	$< 10^{-6}$
$\lambda_{\text{anchor}} \times 10$	8.73	55–62%	0.9995	0.99/0.96	0/9	$< 10^{-6}$
Parameterization and fitting protocol						
No re-orthonormalization	8.94	52–60%	0.995	0.88/0.61	1/9	0.42
Fit jointly with backbone	8.75	48–58%	0.99	0.90/0.70	0/9	$< 10^{-6}$
G attached in the projection	8.72	51–58%	0.999	0.93/0.80	0/9	$< 10^{-6}$
Control						
Random orthonormal G, unanchored	9.24	49–51%	n/a	0.06/0.01	4/9	$< 10^{-6}$

near the cone rather than in it and lets rows drift toward each other so that a constraint on one channel reaches another. Fourth, the detachment of G inside the projection is a precaution rather than a repair: with G attached, a rotation that made the constraint vacuous is available as a descent direction, but it is not taken, and the binding rate stays at 51 to 58%. We keep the detachment because nothing in the objective rules the rotation out, not because we observed it.

The Control row replaces G with a random orthonormal frame and fits nothing. Its rows agree across seeds at chance, four of the nine scored channels reverse, and OOD error rises to 9.24, above the default at 8.69. A cone imposed in arbitrary directions is worse than the anchored one.

D.2 How far back the second pass reaches

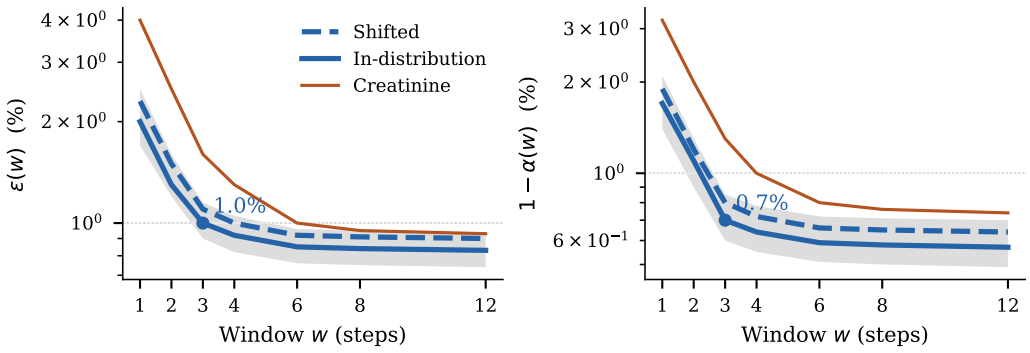


Figure 5: Effect of the trailing window size w on relative state change error $\varepsilon(w)$ (left) and decision inconsistency $1 - \alpha(w)$ (right). Solid curves denote in-distribution slices, dashed curves denote shifted slices, and the orange curve represents slow channels (e.g., creatinine). The chosen default window $w = 3$ is marked with a dot.

Section 4 withdraws every recorded dose, so Δz_{t+1} carries the contribution of the whole dose history, while the cone is built from the drugs active at t. To check that the two refer to nearly the same interventions, let $\Delta z_{t+1}^{(w)}$ be the difference obtained by withdrawing doses only over the trailing w

Table 9: Ablation of PHYSIOWORLD on MIMIC-IV. Each row removes a single component from the full method; rows are *not* cumulative. Mean over five seeds.

Configuration	Skill (↑)	RMSE (↓)
PHYSIOWORLD	**+0.20**	**8.69**
− sign projection	+0.11	9.68
− second forward pass	+0.08	10.08
Backbone (JEPA)	+0.05	10.45

steps, and write

$$\varepsilon(w) \;=\; \mathrm{median}_t \;\frac{\left\|\Delta z^{(w)}_{t+1} - \Delta z_{t+1}\right\|}{\left\|\Delta z_{t+1}\right\|},$$

$$\alpha(w) \;=\; \Pr_{(t,c):\,\sigma_c \text{ defined}}\Big[\mathrm{sign}\big(G\Delta z^{(w)}\big)_c = \mathrm{sign}\big(G\Delta z\big)_c\Big].$$

ε measures how much of the effect a short window misses; α measures how often that miss would change a clipping decision, which is the only thing the operator reads. We find $\varepsilon(3) = 1\%$ and $\alpha(3) = 99.3\%$, with Figure 5 giving both against w per channel and separately on the in-distribution and shifted slices. Because slow channels decay more gradually over time, the cone acts as a conservative approximation for them.

D.3 WHAT THE SECOND PASS WITHDRAWS

The two forward passes described in Section 4 differ in drug doses alone, since the ventilator setpoints and the dialysis and ventilation status flags share the intervention tensor but carry no dose semantics, and withdrawing them would ask what happens with the ventilator switched off. Their contribution would then enter Δz and be projected, although the corpus asserts nothing about them.

D.4 WHY THE INTERVENTION STREAM STAYS IN THE RECURRENT CORE

An alternative is to sever the intervention stream from the recurrent core, in which case setting $u = 0$ would return the untreated course exactly, but we do not do this, because a patient's autonomous course depends on their intervention history: without that history, the core would lose information about the patient's state, for example that norepinephrine has been running for six hours.

E COMPONENT ABLATION

With the corpus held fixed, we remove one component at a time and report out-of-distribution skill and RMSE, averaged across all five shift axes (Table 9). Removing the sign projection while keeping the two-pass decomposition tests whether the decomposition alone already regularizes the predictor: skill falls to $+0.11$, retaining 40% of the gain over the backbone, so the decomposition carries signal on its own and the projection accounts for the rest. Removing the second forward pass and applying the projection to the decoded prediction $\hat{y}$ instead of the isolated effect $\Delta z = z(u) - z(0)$ is more damaging: skill drops to $+0.08$, within 0.03 of the unconstrained backbone, so the projection helps mainly when it acts on the intervention effect itself. Since neither component alone recovers the full method, the gain comes from their combination: the second pass isolates a quantity about which pharmacology makes statements, and the projection enforces them on it.

F HYPERPARAMETERS AND TRAINING DETAILS

Table 10 lists all settings used for the reported runs; settings not listed are left at their defaults. Epoch counts are the same for every slice, and no minimum number of gradient steps is imposed on smaller slices. The operator's calibration map G is a $C \times d$ matrix with orthonormal rows, initialized as a random orthonormal frame and trained only by an anchor loss, which reads the latent out onto the physical channels. The projection uses a detached copy of G, so the constraint does not

Table 10: Hyperparameters. PHYSIOWORLD and every baseline share the data interface and the optimization settings; the two stages differ only in what they train and in how the model is selected.

Setting	Value	Setting	Value
Data interface		*Stage two: readout, backbone frozen*	
Grid resolution	1 h	Readout	MLP $d \to d_{\text{model}} \to C$
Horizon per stay	$\leq$ 96 h or 512 steps	Epochs	15
Batch size	16	Optimizer	AdamW, lr 10^{-3}
Splits	subject-disjoint	Model selection	validation RMSE
Seeds	5		
Backbone		*Shared*	
Hidden width d_{model}	128	Weight decay	0
GRU layers	3	Gradient clipping	max-norm 1.0
Latent dimension d	32	Non-finite batches	skipped, not stepped
Dropout	0.1		
Intervention encoder	linear		
Stage one: encoder, predictor, intervention pathway		*Baselines*	
Epochs	8	General	published defaults
Optimizer	AdamW, lr 10^{-3}	Windowed forecasters	lookback 24 steps
Schedule	cosine decay to 10^{-5}		
SIGReg weight λ	0.05		
Auxiliary loss weight	0.5		
Model selection	validation latent loss		

shape the map through which it is expressed, and since the condition $GG^\top = I$ keeps the write-back minimum-norm and confined to G's row space but is not preserved by a gradient step, G is re-orthonormalized after every optimizer step.

G HOLDOUT CONSTRUCTION

Every partition follows the split-then-slice protocol from Setup, with a deterministic seed for the subject-level split, and a model trained on one slice's training set is evaluated on the test set of every slice in the same partition, where the entries whose test slice differs from the training slice give the out-of-distribution error. This section lists the strata.

Population axes. *Sex* produces two slices (male, female), and *Age* uses the five bins of BEDS-Bench (Avati et al., 2021), a published benchmark for evaluating EHR models under distribution shift: $(15, 50]$, $(50, 60]$, $(60, 70]$, $(70, 80]$, and $80+$. *Care unit* takes the first ICU unit of each hospital admission (resolving transfers), keeping at most the four largest units by admission count, while *Disease subtype* maps the primary discharge diagnosis through ICD-9/10 ranges into ten circulatory-system groups (rheumatic and valvular, hypertensive, ischemic heart disease, pulmonary heart disease, heart failure, arrhythmia, cardiomyopathy and other heart disease, cerebrovascular, peripheral vascular, venous and other), keeping at most the four largest and assigning no disease slice to a stay whose diagnosis falls outside the cohort's ICD chapter. *Era* uses MIMIC-IV's `anchor_year_group` field, which bins each patient's first admission into a four-year window (2008–2010, 2011–2013, 2014–2016, 2017–2019), and all era slices are kept.

Intervention-intensity axes. Four drug-class axes (vasopressor, sedative, rate control, vasodilator) stratify stays by how intensively the patient was treated, and for each class a per-stay *burden* is computed: each drug's per-stay peak dose is replaced by its within-drug rank among exposed stays, and the class burden is the maximum rank across drugs in the class, which avoids any assumption of cross-drug equipotency. Stays with zero exposure to the class receive no burden score and join no slice, and exposure to the vasopressor and sedative classes is 40.9% and 64.0% of stays on MIMIC-IV circulatory and 39.6% and 44.9% on MIMIC-III circulatory. Among exposed stays, the 25th and 75th percentiles of the burden define three bands, *low* ($\leq$ P25), *mid* (P25–P75), and *high* ($>$ P75), of which only the low and high bands become slices, the mid band serving as a buffer that keeps patients near a cut-off out of both extremes. An intervention axis therefore contributes two slices wherever both bands clear the minimum slice size below, and fewer where they do not.

Table 11: Slices per partition on MIMIC-IV circulatory, in admissions. A partition's excluded count is the number of its groups that fall below the minimum slice size.

Partition and slice	Train	Test	Partition and slice	Train	Test
Sex			*Era*		
female	6162	1536	2008–2010	4532	1148
male	8824	2233	2011–2013	3407	863
Age			2014–2016	3705	944
15–50	1702	412	2017–2019	3342	814
50–60	2611	685	*Vasopressor dose*		
60–70	4061	972	high	1940	490
70–80	3899	968	low	1903	492
80+	2713	732	*Sedative dose*		
Care unit (5 excluded)			high	2336	560
CVICU	6459	1592	low	2227	564
CCU	3165	765	*Rate-control dose*		
MICU	1158	320	high	893	224
SICU	1801	436	low	1321	335
Disease subtype (6 excluded)			*Vasodilator dose*		
cardiomyopathy / other heart	1536	380	high	1709	438
cerebrovascular	3456	880	low	1703	404
ischemic heart disease	4501	1152			
rheumatic / valvular	1180	304			

The four drug classes and their constituent drugs are:

- *Vasopressors*: norepinephrine, epinephrine, phenylephrine, vasopressin, dopamine, dobutamine, milrinone.
- *Sedatives*: propofol, fentanyl, midazolam, dexmedetomidine, lorazepam, diazepam.
- *Rate control*: metoprolol, esmolol, labetalol, amiodarone. On the CareVue-era MIMIC-III database, metoprolol and labetalol lack canonical charting identifiers, leaving only esmolol and amiodarone, which reduces the exposure rate.
- *Vasodilators*: nitroglycerin, nicardipine, nitroprusside, hydralazine.

Minimum slice size. Any slice with fewer than 450 distinct hospital admissions (across train and test combined) is excluded from its partition. This threshold is a run parameter and is set lower than BEDS-Bench's default of 2 000 because the diagnosis-selected cohorts are roughly half the size of the full MIMIC population.

Training-set balancing. Within each partition, every slice's training set is downsampled to the admission count of the smallest slice (sampling without replacement at the admission level, re-drawn per seed), so that cross-slice comparisons are not confounded by unequal training volume. Table 11 lists the resulting slices on MIMIC-IV circulatory, where counts are distinct hospital admissions at seed 0 and a test fraction of 0.2, before training-set balancing; after balancing, every slice in a partition trains on the admission count of the smallest slice in that partition.

Not every axis is available on every cohort, and Table 12 gives the number of slices each partition yields. Era is a MIMIC-IV field and has no counterpart in the CareVue era, VitalDB charts no rate-control or vasodilator agent, so those two axes cannot be built on it, and a partition needs at least two slices to contribute an off-diagonal, which rules out rate control on MIMIC-III circulatory and the vasodilator axis on MIMIC-IV injury.

H Cost

PhysioWorld evaluates the transition dynamics with and without the administered intervention, which requires two forward passes per step. Measured on a single device under a fixed batch and sequence configuration, a training step for PhysioWorld takes 46.9 ms against 21.0 ms for the unconstrained JEPA backbone (Table 13), a factor of 2.23 that slightly exceeds the 2.0 expected from the two passes, with the remainder attributable to the projection. The projection itself is parameter-free, and the only added weights belong to the calibration map G ($C \times d = 288$ parameters), 0.07% of the model's parameters, while inference takes 25.2 ms per step against the backbone's 6.6 ms. For

Table 12: Slices per partition on each cohort, at the same minimum slice size. A dash is a partition that yields no slice, because its groups all fall below the minimum size or the field does not exist in that database; *n/a* is a partition whose drug class is never charted in that database.

	Population					*Intervention intensity*			
Cohort	Sex	Age	Care unit	Disease	Era	Vasopr.	Sedative	Rate ctrl.	Vasodil.
MIMIC-IV circ.	2	5	4	4	4	2	2	2	2
MIMIC-IV inj.	2	5	4	4	4	2	2	2	1
MIMIC-III circ.	2	5	4	4	—	2	2	1	2
MIMIC-III inj.	2	5	3	2	—	—	2	—	—
VitalDB	2	4	2	—	—	—	2	n/a	n/a

Table 13: Cost of one step, for every method in Table 1. Median over many steps after warm-up, one device, fixed batch size and sequence length, identical data for every architecture, with every timed region synchronized. Chronos-2 is zero-shot, so it has no training step to time. PatchTST's attention runs over patches, so its cost is $\mathcal{O}(T^2 d/P^2)$ in a patch length P.

Method	Train ms	Infer ms	Peak MB	Params
Reference				
Persistence	2.6	0.2	1.4	20 K
Chronos-2	—	7 839	338	119.5 M
Recurrent: $\mathcal{O}(Td^2)$, linear in T				
CRN	30.8	16.5	87	341 K
IRM	35.7	6.4	173	402 K
CORAL	58.7	6.4	174	402 K
JEPA (backbone)	21.0	6.6	173	402 K
Soft constraint	75.0	6.5	174	402 K
Clamp	43.4	19.5	196	375 K
PHYSIOWORLD	**46.9**	**25.2**	**199**	**402 K**
Attention: $\mathcal{O}(T^2 d)$, quadratic in T				
Causal Transformer	72.6	23.5	238	1.9 M
PatchTST	75.5	58.0	1 188	88 K
GraFITi	146.9	160.2	2 619	657 K

clinical deployment, a more relevant comparison is with a pretrained forecaster: Chronos-2 needs 7 839 ms per inference step and 119.5 M parameters, about 300 times the runtime and the parameter count of PHYSIOWORLD. The overhead is a constant factor, and PHYSIOWORLD keeps the linear $\mathcal{O}(Td^2)$ complexity and memory footprint of recurrent architectures.

I CONVERGENCE

To check that no model is undertrained under the budget of Appendix F, we record validation loss at the end of every epoch, and Figure 6 shows the curves. The gap between the final and the best epoch is not informative here, because it is zero in nearly every run and is satisfied by construction, so we instead report how much improvement the last quarter of the budget still yields, which is between 0.01 and 0.45%, against total improvements of 0.9 to 11.9%. Every architecture has reached a plateau when the budget ends.

J PROOF OF PROPOSITION 1

Proof. Let $\mathcal{I}$ be the set of coordinates on which σ_c is defined. Then $\mathcal{C} = \prod_c C_c$, where $C_c = \{t \in \mathbb{R} : \sigma_c t \geq 0\}$ for $c \in \mathcal{I}$ and $C_c = \mathbb{R}$ otherwise. Each C_c is a closed convex cone, so $\mathcal{C}$ is a closed convex cone and is nonempty, since $0 \in \mathcal{C}$. Because $\mathcal{C}$ is a Cartesian product, the Euclidean projection onto it separates across coordinates, and on a half-line the nearest point to t is $\sigma_c \max(\sigma_c t, 0)$; the map Π of equation 3 is therefore the Euclidean projection onto $\mathcal{C}$.

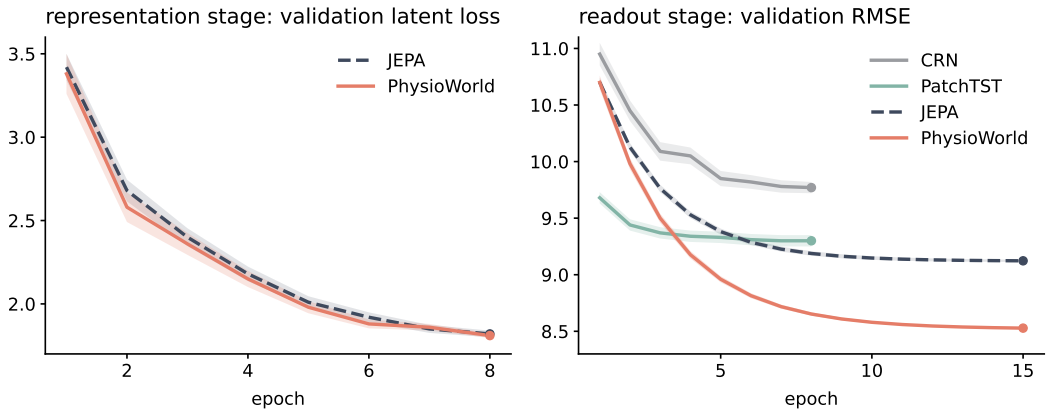


Figure 6: Validation loss across epochs on MIMIC-IV circulatory, mean over five seeds with ± 1 std shaded and the best epoch marked. CRN and PatchTST train in one stage; the JEPA family trains a representation against a latent loss and then a readout on the frozen representation, so its right-hand curve begins near its floor. Every curve flattens before the budget ends.

Claims (i) and (ii) are then immediate. For (iii), each coordinate map is the identity, $\max(\cdot, 0)$, or $\min(\cdot, 0)$, and each is 1-Lipschitz on $\mathbb{R}$; hence $|\Pi(y)_c - \Pi(y')_c| \le |y_c - y'_c|$ for every c, and summing squares gives the claim. For (iv), the projection theorem on a closed convex set gives $\langle y - \Pi(y),\, w - \Pi(y)\rangle \le 0$ for every $w \in \mathcal{C}$; taking $w = y^\star$ yields $\langle y - \Pi(y),\, \Pi(y) - y^\star\rangle \ge 0$. Decomposing $y - y^\star = \big(y - \Pi(y)\big) + \big(\Pi(y) - y^\star\big)$ and expanding,

$$\|y - y^\star\|^2 - \|\Pi(y) - y^\star\|^2 \;=\; \|y - \Pi(y)\|^2 + 2\big\langle y - \Pi(y),\, \Pi(y) - y^\star\big\rangle \;\ge\; \|y - \Pi(y)\|^2. \quad \square$$

K Significance against Clamp

We compare PhysioWorld with Clamp, the strongest method sharing our backbone, using two tests whose assumptions differ. The first is a patient-level paired bootstrap (Table 14): for each axis and cohort, test patients are resampled with replacement 1,000 times, both methods are scored on the same resample, and the 2.5th and 97.5th percentiles of the RMSE difference give a 95% interval. Resampling patients rather than steps respects the dependence between steps of the same stay. The second is a one-sided Wilcoxon signed-rank test over the 56 train/test slice pairs, 48 from the five axes of MIMIC-IV circulatory and 8 from the sedative axis of the other four cohorts, with seeds averaged within each pair. Slice pairs on the same axis share their training slices and are not independent, so the Wilcoxon p-value is approximate; the bootstrap makes no assumption across slice pairs and is the primary test.

Table 14: RMSE difference Clamp $-$ PhysioWorld with patient-level 95% bootstrap intervals; positive favors PhysioWorld.

Cohort	Axis	ΔRMSE	95% interval	Excludes 0
MIMIC-IV circ.	Age	0.14	$[0.09, 0.19]$	yes
	Disease	0.14	$[0.08, 0.20]$	yes
	Era	0.05	$[-0.01, 0.11]$	no
	Vasopressor	0.23	$[0.12, 0.34]$	yes
	Sedative	0.17	$[0.07, 0.27]$	yes
MIMIC-IV injury	Sedative	0.09	$[-0.02, 0.20]$	no
MIMIC-III circ.	Sedative	0.08	$[-0.03, 0.19]$	no
MIMIC-III injury	Sedative	0.06	$[-0.08, 0.20]$	no
VitalDB	Sedative	0.07	$[-0.06, 0.20]$	no

Across the 56 slice pairs, PhysioWorld has the lower error on 48, with a median difference of 0.11, and the one-sided Wilcoxon test gives $p < 10^{-6}$.

L ROLLOUT AT EVERY HORIZON

Table 15 gives the autoregressive RMSE of Section 5.3 at every horizon with its across-seed standard deviation, together with the dispersion at the first and last step.

Table 15: Free-running autoregressive rollout on MIMIC-IV after a half-stay warmup, mean and standard deviation over five seeds. Slope is the least-squares fit of $\log \mathrm{RMSE}$ on $\log h$ over $h \in \{1, 2, 4, 8\}$; dispersion is $\mathrm{sd}_{\mathrm{pred}}/\mathrm{sd}_{\mathrm{truth}}$ at h=1 and h=8.

	Autoregressive RMSE (↓)					**Dispersion (→ 1)**	
Method	h=1	h=2	h=4	h=8	Slope (↓)	h=1	h=8
Reference and data-driven forecasters							
Persistence	11.48±0.03	12.18±0.04	12.89±0.04	13.66±0.05	+0.084	1.000	0.985
PatchTST	8.89±0.02	9.68±0.03	10.58±0.03	11.58±0.04	+0.127	0.884	0.631
GraFITi	9.11±0.02	9.85±0.03	10.68±0.03	11.58±0.04	+0.116	0.892	0.700
Chronos-2	8.85±0.02	9.60±0.03	10.45±0.03	11.42±0.04	+0.124	0.880	0.645
CRN	9.85±0.03	10.58±0.03	11.42±0.04	12.31±0.05	+0.108	0.871	0.758
Causal Transformer	9.28±0.02	10.02±0.03	10.85±0.03	11.78±0.04	+0.115	0.865	0.663
Robustness by reweighting or invariance							
IPTW	9.35±0.03	10.82±0.06	13.60±0.14	17.15±0.25	+0.296	0.905	1.825
IRM	9.39±0.03	12.50±0.05	17.80±0.12	24.85±0.22	+0.472	0.912	1.757
CORAL	10.21±0.03	11.15±0.04	11.92±0.04	12.65±0.05	+0.102	0.612	0.415
World models							
JEPA	9.12±0.02	9.95±0.03	11.10±0.04	12.45±0.05	+0.150	0.895	0.748
Soft constraint	8.85±0.02	9.62±0.03	10.65±0.04	11.85±0.04	+0.141	0.902	0.765
Clamp	8.92±0.02	9.58±0.02	10.42±0.03	11.35±0.04	+0.116	0.910	0.792
PHYSIOWORLD	**8.52**±0.01	**9.04**±0.02	**9.62**±0.02	**10.28**±0.03	**+0.090**	**0.935**	**0.842**