# BEYOND PATIENT INVARIANCE: LEARNING CARDIAC DYNAMICS VIA ACTION-CONDITIONED JEPAS

**Jose Geraldo Fernandes, Luiz Facury de Souza, Pedro Robles Dutenhefner & Wagner Meira Jr.**
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{josefernandes,luizfacury,pedroroblesduten,meira}@dcc.ufmg.br

## ABSTRACT

Self-supervised learning in healthcare has largely relied on invariance-based objectives, which maximize similarity between different views of the same patient. While effective for static anatomy, this paradigm is fundamentally misaligned with clinical diagnosis, as it mathematically compels the model to suppress the transient pathological changes it is intended to detect. We propose a shift towards Action-Conditioned World Models that learn to simulate the dynamics of disease progression, or *Event-Conditioned*. Adapting the LeJEPA framework to physiological time-series, we define pathology not as a static label, but as a transition vector acting on a patient's latent state. By predicting the future electrophysiological state of the heart given a disease onset, our model explicitly disentangles stable anatomical features from dynamic pathological forces. Evaluated on the MIMIC-IV-ECG dataset, our approach outperforms fully supervised baselines on the critical triage task. Crucially, we demonstrate superior sample efficiency: in low-resource regimes, our world model outperforms supervised learning by over 0.05 AUROC. These results suggest that modeling biological dynamics provides a dense supervision signal that is far more robust than static classification. Source code is available at our GitHub repository.

## 1 INTRODUCTION

The primary goal of clinical monitoring is the detection of change. Whether observing a patient in an intensive care unit or tracking chronic conditions over years of outpatient visits, the clinical value of a diagnostic tool lies in its ability to distinguish the stable anatomical baseline of an individual from acute pathological transitions. However, the current landscape of self-supervised learning (SSL) in medical artificial intelligence is governed by a fundamental contradiction: it is built almost entirely on the principle of *invariance*.

In standard SSL frameworks, such as SimCLR or MoCo, and their medical adaptations like PCLR (Diamant et al., 2022) and MICLe (Azizi et al., 2021), the objective is to learn representations that remain stable under perturbations. In the context of longitudinal data, this often manifests as *Patient-Invariance*, where a model is trained to maximize the similarity between different recordings of the same individual. While effective for tasks involving static anatomical characterization or patient re-identification, we argue that this paradigm is inherently misaligned with the requirements of acute care. By forcing a model to map a healthy baseline and an acute pathological state (e.g., the onset of a paroxysmal arrhythmia) to the same point in a latent manifold, the encoder is mathematically compelled to discard the transient disease signals as noise or augmentation artifacts (Mehari & Strodthoff, 2022). This creates a fundamental information bottleneck: a robust biometric model that is paradoxically blind to the very changes it is meant to monitor.

Furthermore, the standard data augmentations used to provide the *views* for contrastive learning—such as random cropping, jittering, or scaling—often fail when applied to 1D physiological signals like the Electrocardiogram (ECG). Techniques like random cropping can be semantically destructive, potentially removing the P-wave and rendering a model unable to assess the P-R interval (Andersson, 2023), while jittering and scaling can inadvertently mimic serious pathologies (Non-

nenmacher et al., 2022). Consequently, a model trained to be invariant to these transformations may unlearn critical clinical markers essential for diagnosis.

To resolve these limitations, we propose a shift from static invariance to *dynamic causality*. We introduce the Action-Conditioned Cardiac World Model, a framework that formalizes the monitoring problem as a simulation of the heart's electrophysiological state under pathological perturbations. Rather than suppressing longitudinal variations, our model explicitly treats disease onset as an *action*—a translational force that moves the patient's representation through a latent manifold. By utilizing the Joint-Embedding Predictive Architecture (JEPA) (LeCun, 2022), we avoid the computational burden and "hallucination" risks associated with generative pixel-reconstruction models like MeWM (Yang et al., 2025). Instead, we predict future states directly in a geometrically regularized latent space.

This approach allows the model to disentangle the stable anatomical substrate (Patient Identity) from the evolving disease physics. Our experiments on the MIMIC-IV-ECG dataset demonstrate that this dynamical systems perspective achieves parity with fully supervised models on longitudinal monitoring while significantly outperforming them on the critical task of triage for new patients. Most importantly, we show that dynamics modeling provides a denser supervision signal than classification, maintaining high performance in low-data regimes where supervised baselines suffer catastrophic degradation.

Our contributions are as follows:

- We propose an **Action-Conditioned World Model** for medical diagnosis, recontextualizing disease onset as a translational action variable.
- We demonstrate that **predictive dynamics serve as a superior supervision signal** to standard classification, providing a 0.05 AUROC improvement in low-resource (10%) data regimes.
- We show that our model **overcomes the limitations of standard data augmentations** by using longitudinal transitions as the primary source of supervision, resulting in more robust representations for clinical monitoring.

## 2 Related Work

**Invariance-based Self-Supervised Learning.** The dominant paradigm in medical representation learning has centered on the principle of invariance. Approaches such as Patient Contrastive Learning of Representations (PCLR) (Diamant et al., 2022) maximize similarity between different recordings of the same individual to learn robust anatomical features. While successful for static tasks like sex classification or age regression, this patient-invariance creates a fundamental information bottleneck for clinical monitoring. As demonstrated by the observed *AFib Gap*, forcing a model to map a healthy baseline and an acute pathological state (e.g., paroxysmal Atrial Fibrillation) to the same latent point compels the encoder to discard the very disease signals it is intended to detect. Similar limitations are observed in Multi-Instance Contrastive Learning (MICLe) (Azizi et al., 2021), which excels in static domains like dermatology but collapses when applied to transient pathologies. Furthermore, efforts to remove subject variability in EEG (Salsabilian & Najafizadeh, 2022) often result in a loss of sensitivity to personalized anomalies, such as pre-ictal states in epilepsy. Our work diverges from this paradigm by treating patient identity not as a nuisance to be invariant to, but as a stable substrate that must be disentangled from dynamic pathological transitions.

**Medical World Models.** The concept of "World Models"—internal simulations that predict the consequences of actions—has recently transitioned from robotics to healthcare. Generative World Models, such as MeWM (Yang et al., 2025), simulate tumor evolution conditioned on treatment plans but rely on high-fidelity pixel reconstruction, which is computationally expensive and prone to hallucinations that lack biological reality. More recent efforts like SurgWM (Shah et al., 2026) utilize latent dynamics to *imagine* surgical tissue deformation based on tool movement, demonstrating the utility of predicting in embedding space. In the context of static imaging, CheXWorld (Yue et al., 2025) treats spatial exploration as a dynamic process, using a JEPA framework to disentangle anatomy from domain variation. Our approach extends this logic to the temporal domain of ECG,

recontextualizing disease onset not as a static label, but as a translational *action* on the patient's latent identity.

**Latent Dynamics and JEPA in Time-Series.** Our architecture builds upon the Joint-Embedding Predictive Architecture (JEPA) (LeCun, 2022), which avoids the pitfalls of both generative reconstruction and contrastive invariance. While Masked Autoencoders (MAE) approaches like Zhou et al. (2025) focus on high-frequency texture and sensor noise (Zhang et al., 2024), JEPA-style predictive objectives (Ren et al., 2022) capture semantic causal structures by predicting future latent states. JETS (Xie et al., 2025) represents a primary application of JEPA to physiological time-series; however, it remains purely autoregressive (passive). Our model addresses this action gap by injecting a pathology transition vector, upgrading the model from a passive forecaster to an active simulator of cardiac evolution. This transition is further supported by literature on time-irreversibility (Agrawal et al., 2022), which argues that because disease progression is entropic and directed, models must respect the "arrow of time" rather than treating longitudinal pairs as interchangeable views.

**Disease Progression as a Dynamical System.** Modeling pathology as a trajectory through a latent manifold aligns with recent advances in Neural Ordinary Differential Equations (NODEs). LN-ODE (Wen & Biros, 2025) demonstrates the power of disentangling subject-specific parameters (identity) from cohort-shared parameters (disease physics) to improve forecasting. While previous longitudinal SSL methods (LSSL) (Zhao et al., 2021) regularize latent space based on temporal smoothness, they often struggle with the discontinuous jumps caused by acute events like myocardial infarction. By conditioning on an explicit action vector, our framework linearizes these complex differential equations, allowing for the simulation of both gradual progression and sudden pathological onsets. This bridges the gap between self-supervised representation learning and causal counterfactual regression (El Bouchattaoui et al., 2024), enabling the model to answer "what-if" questions regarding a patient's cardiac future.

**Limitations of Data Augmentation in ECG.** Finally, our move toward predictive dynamics is motivated by the failure of standard SSL augmentations for periodic signals. Common techniques like random cropping can be semantically destructive in ECG, potentially removing the P-wave and rendering the model unable to assess the P-R interval (Andersson, 2023). As noted in recent critiques of ECG augmentation (Mehari & Strodthoff, 2022; Nonnenmacher et al., 2022), standard transformations like jittering and scaling risk introducing invariances to features that are morphologically similar to acute pathologies. Mehari & Strodthoff (2022) demonstrated that such artificial views can lead to significantly reduced robustness compared to supervised baselines, effectively causing the model to unlearn critical clinical markers. By leveraging longitudinal transitions as the primary supervision signal, our model avoids the need for ad-hoc heuristics that risk unlearning critical clinical features.

## 3 METHOD: ACTION-CONDITIONED CARDIAC WORLD MODELS

We formalize the cardiac monitoring problem as learning a world model that simulates the evolution of the heart's electrophysiological state under pathological perturbations. Unlike standard SSL which enforces invariance to augmentation, our objective is to learn a predictive latent space where disease onset acts as a translational force on the patient's anatomical representation, leveraging patient identity information. The architecture is schematized as in Figure 1.

**Problem Formulation.** Let $\mathbb{X}$ be the space of 12-lead ECG signals $\boldsymbol{X} \in \mathbb{R}^{12\times T}$ and $\mathbb{Y}$ be the associated set of diagnosis labels $\boldsymbol{y} \in \{0,1\}^C$. In longitudinal datasets (e.g., MIMIC-IV), we observe a sequence of states $(\boldsymbol{X}_t, \boldsymbol{y}_t)$ for a specific patient.

We define the action $\boldsymbol{a}_t$ not as the static label, but as the transition vector describing the pathological change between time steps:

$$\boldsymbol{a}_t = \boldsymbol{y}_{t+1} - \boldsymbol{y}_t \in \{-1, 0, 1\}^C$$

Here, $\boldsymbol{a}_{t,i} = 1$ denotes disease onset, $-1$ denotes resolution, and $0$ implies stability. The goal of the World Model is to learn an encoder $\text{Enc}_\theta : \mathbb{X} \to \mathbb{H}$ and a predictor $\text{Dyn}_\phi$ such that the future latent state can be estimated from the current state and the applied clinical *action*, after a projection

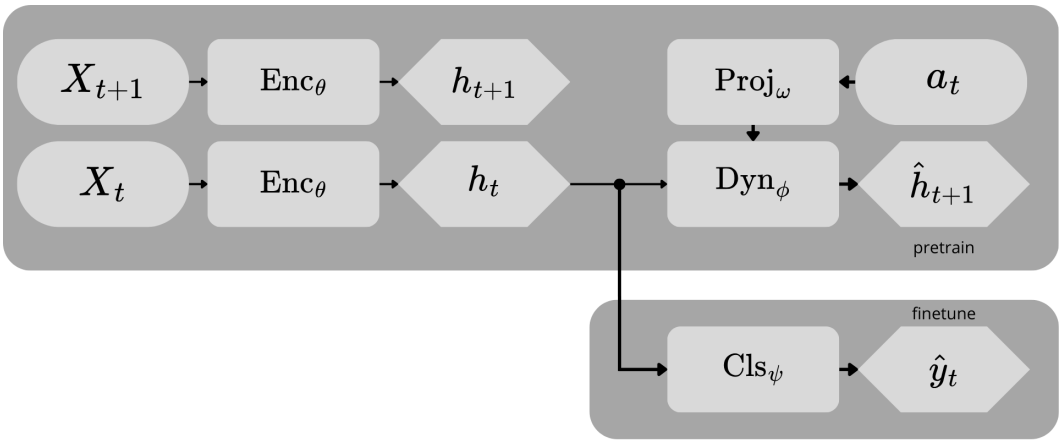


Figure 1: Proposed Action-Conditioned World Model Architecture. The framework consists of two stages. Top (Pre-training): the encoder $\text{Enc}_\theta$ maps longitudinal ECG pairs $(\boldsymbol{X}_t, \boldsymbol{X}_{t+1})$ to latent representations $(\boldsymbol{h}_t, \boldsymbol{h}_{t+1})$. A dynamics network $\text{Dyn}_\phi$ predicts the future latent state $\hat{\boldsymbol{h}}_{t+1}$ conditioned on the current state $\boldsymbol{h}_t$ and the pathology transition vector $\boldsymbol{a}_t = \boldsymbol{y}_{t+1} - \boldsymbol{y}_t$, processed by projector $\text{Proj}_\omega$. Bottom (Fine-tuning): the learned representation $\boldsymbol{h}_t$ is passed to a classifier $\text{Cls}_\psi$ to predict multi-label diagnoses $\hat{\boldsymbol{y}}_t$.

via $\text{Proj}_\omega$: $\hat{\boldsymbol{h}}_{t+1} = \text{Dyn}_\phi(\text{Enc}_\theta(\boldsymbol{X}_t), \text{Proj}_\omega(\boldsymbol{a}_t))$. This formulation forces the encoder $\text{Enc}_\theta$ to disentangle Patient Identity (invariant features required to ground the prediction) from Disease State (dynamic features modified by $\boldsymbol{a}_t$).

**Dynamics Visualization.** To illustrate the transition vector's role in grounding the latent space, we visualize specific longitudinal ECG pairs $(\boldsymbol{X}_t, \boldsymbol{X}_{t+1})$. In Figure 2, the action $\boldsymbol{a}_t$ is active only for label I44, which is morphologically represented by the wide QRS complex (black arrow). In Figure 3, while $\boldsymbol{a}_t$ remains still active for I44, both signals share a common stable label in $\boldsymbol{y}$ (I48), consequently, this component is transparent in the action vector $\boldsymbol{a}_{t,\text{I48}} = \boldsymbol{y}_{t+1,\text{I48}} - \boldsymbol{y}_{t,\text{I48}} = 0$, distinct from the flutter waves (red arrow) visible in the traces. Critically, these examples illustrate the disentanglement objective, the model is trained to linearize disease progression, isolating the transition vector as an independent force acting upon a stable, patient-specific substrate.

**Architecture.** We adopt the JEPA framework, adapted for 1D physiological signals. The system consists of an encoder $\text{Enc}_\theta$ implemented as an xResNet1d50, a 1D-adapted ResNet architecture optimized for time-series following previous ECG classification literature (Strodthoff et al., 2024b), which maps the raw waveform $\boldsymbol{X}_t$ to a latent representation $\boldsymbol{h}_t$. Simultaneously, the sparse difference vector $a_t$ is projected into a continuous action embedding via a multi-layer perceptron (MLP) as a projector $\text{Proj}_\omega$. The Latent Dynamics Predictor $\text{Dyn}_\phi$, a residual MLP, then estimates the future representation $\hat{\boldsymbol{h}}_{t+1} = \text{Dyn}_\phi(\text{Enc}_\theta(\boldsymbol{X}_t), \text{Proj}_\omega(\boldsymbol{a}_t))$ based on the current state and given the action. Note that unlike generative models, we predict in latent space, avoiding the complexity of pixel-level reconstruction.

**Regularization via SIGReg.** A critical failure mode in predictive SSL is representation collapse (mapping all inputs to a constant vector). While methods like VICReg (Bardes et al., 2021) prevent this via heuristic variance-covariance constraints, they introduce brittle hyperparameters that are difficult to tune. We instead employ Sketched Isotropic Gaussian Regularization (SIGReg) from the LeJEPA framework. SIGReg enforces that the distribution of embeddings $\mathbb{H}$ follows a standard Isotropic Gaussian $\mathcal{N}(\boldsymbol{h}; \boldsymbol{0}, \boldsymbol{I})$ by minimizing the distance between their empirical characteristic functions projected onto random 1D slices.

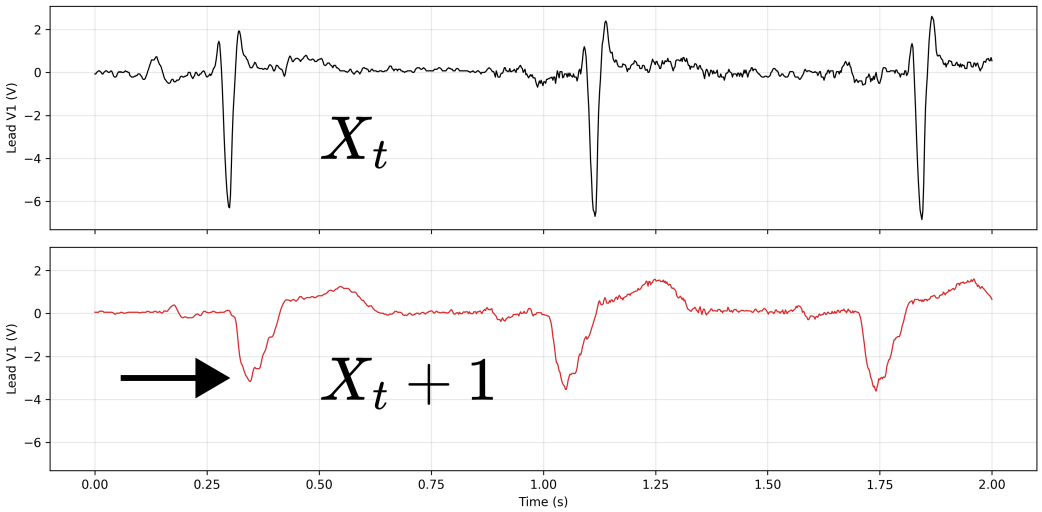


Figure 2: Visualization of a Pure Pathological Transition. This longitudinal pair illustrates a transition where the action vector $\boldsymbol{a}_t$ is sparse, representing the onset of a single pathology (I44). The baseline signal (top) shows a normal ventricular activation, while the subsequent recording (bottom) exhibits the classic morphological hallmark of I44, a widened QRS complex (black arrow). By conditioning on this transition vector, the world model is trained to simulate the transformation from a healthy baseline to this pathological state within the latent manifold.

The total training objective is:

$$\mathcal{L} = (1-\lambda)\underbrace{\|\hat{\boldsymbol{h}}_{t+1} - \boldsymbol{h}_{t+1}\|^2}_{\text{Prediction (Dynamics)}} + \lambda \underbrace{(\text{SIGReg}(\boldsymbol{H}_t) + \text{SIGReg}(\boldsymbol{H}_{t+1}))}_{\text{Geometry Constraint}}$$

By enforcing an isotropic Gaussian geometry, SIGReg ensures the latent space is maximally entropic (preventing collapse) and rotationally symmetric, which is theoretically optimal for linear separability in downstream tasks. This allows our model to learn rich, manipulable representations of cardiac dynamics without requiring the extensive grid searches associated with contrastive methods.

**Downstream Task.** To assess whether the learned world model captures clinically relevant features, we evaluate the encoder $\text{Enc}_\theta$ on downstream multi-label classification tasks under two distinct regimes: Linear Probing (Geometry Check); and, Finetuning (Adaptability Check). For the former we freeze the encoder and train a linear head $\text{Cls}_\psi$ on top of $\boldsymbol{h}_t$. This tests if the latent space is linearly separable by disease, validating that the SIGReg geometry constraint successfully organized the pathological states. Then, in finetuning, we unfreeze the encoder and update all weights using the supervised classification loss. This tests if the pre-trained initialization places the model in a better optimization basin than random initialization, particularly in low-data regimes.

For both protocols, we utilize the Asymmetric Loss to handle the extreme class imbalance (e.g., Hypertension vs. Rare Arrhythmias) inherent in the dataset.

## 4 Experiments and Discussion

We evaluate our approach on the MIMIC-IV-ECG dataset (Gow et al., 2023) from PhysioNet (Goldberger et al., 2000), containing ∼800k records from ∼160k patients. We utilize the official patient-stratified folds (Strodthoff et al., 2024a) to ensure zero identity leakage between splits[1]. We compare our LeJEPA-based approach against a fully supervised model trained from scratch and a Naive

[1]See Appendix A for a complete dataset characterization.

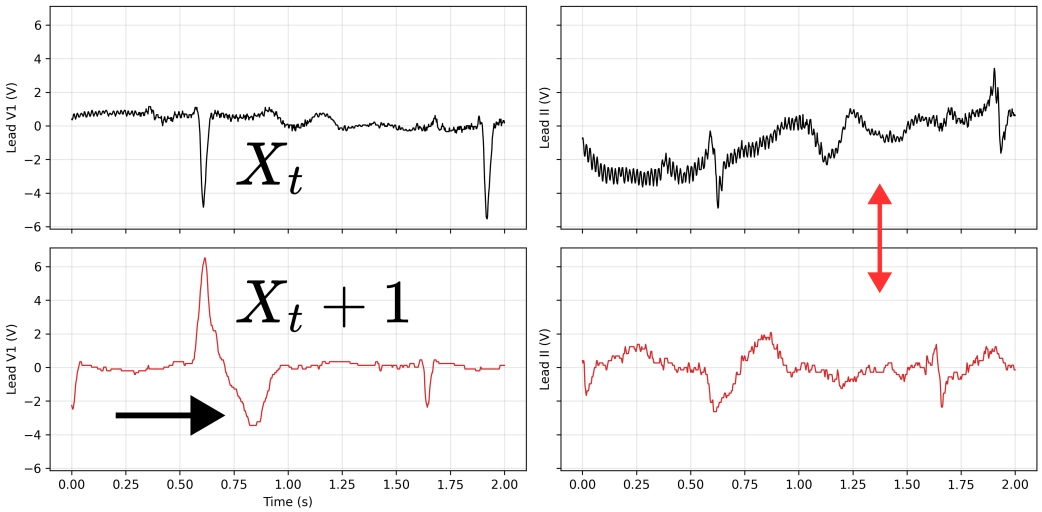


Figure 3: Disentanglement of Acute Pathology from Stable Anatomy. In this more complex clinical scenario, both the baseline and the future state share a stable underlying condition (I48), identified by the persistent flutter waves (red arrow). However, the transition vector $\boldsymbol{a}_t$ again signals the acute onset of I44 (wide QRS complex, black arrow). This demonstrates the model's objective, it must remain invariant to the action-transparent features (flutter waves) while accurately predicting the pathological ones (wide QRS), effectively distinguishing the acute event from the chronic background.

Table 1: Macro-AUROC Performance on MIMIC-IV-ECG. We compare our Action-Conditioned Dynamics model against a supervised baseline and a naive patient-invariance SSL baseline. We report results on the Triage task and the Monitoring task. 95% confidence intervals from 1,000 bootstrap iterations are shown in brackets.

| **Data Usage** | | | | **Macro-AUROC** | |
|---|---|---|---|---|---|
| **Labels ($y$)** | **Patient ID** | **Method** | **Eval. Protocol** | **Triage (First ECG)** | **Monitoring (All)** |
| ✓ | ✗ | Supervised | End-to-End | 0.735 [0.727 – 0.751] | 0.786 [0.776 – 0.797] |
| ✗ | ✓ | Naive SSL (Invariance) | Linear Probe | 0.666 [0.656 – 0.682] | 0.698 [0.689 – 0.709] |
| | | | Finetuned | 0.679 [0.668 – 0.697] | 0.726 [0.715 – 0.737] |
| ✓ | ✓ | Ours (Latent Dynamics) | Linear Probe | 0.725 [0.714 – 0.742] | 0.744 [0.734 – 0.755] |
| | | | Finetuned | 0.742 [0.733 – 0.756] | 0.786 [0.777 – 0.794] |

Patient-Based SSL baseline. We implement the latter by mapping $\mathbf{h}_t \to \mathbf{h}_{t+1}$ without action conditioning, effectively treating longitudinal exams as positive pairs for standard invariance learning.

**Generalization and Robustness.** We first evaluate performance on the full dataset under two distinct clinical scenarios: Triage, evaluating only on the first ECG of a patient's visit; Monitoring, which evaluates on all ECGs but captures the temporal redundancy inherent in multi-record hospital stays. Results are detailed in Table 1.

While both methods achieve parity on the easier Monitoring task, our Dynamics model slightly outperforms the Supervised baseline on the critical Triage task. We hypothesize that the dynamics learning—which requires predicting the transformation of the signal—forces the encoder to disentangle anatomical features from pathological changes, resulting in robust representations. The Naive SSL baseline did not yield good downstream performance for this task, confirming that patient invariance alone is insufficient for diagnosis.

**Data Efficiency.** To test the hypothesis that dynamics modeling provides a denser supervision signal than classification, we simulate low-resource environments by restricting the training data to 10% (∼4k patients) and 1% (∼400 patients) of the dataset. Crucially, unlike standard SSL, we also pretrain with this restriction as our approach requires the label.

As illustrated in Figure 4, the Supervised model suffers catastrophic degradation in the 10% regime ($0.74 \rightarrow 0.64$ AUROC), owing to the high parameter-to-sample ratio. Conversely, our Dynamics model maintains robustness ($0.74 \rightarrow 0.69$ AUROC), significantly outperforming the baseline ($p < 0.05$). This suggests that predicting a high-dimensional future latent state $\boldsymbol{h}_{t+1}$, given the action, provides a richer gradient signal than categorical labels, enabling complex feature learning even when data is scarce. At 1% data, we observe a capability collapse where both models degrade to random guessing (∼0.50 AUROC). This establishes a lower bound for the xResNet1d50 architecture on this modality, indicating that below this threshold, model capacity exceeds the information content of the subset regardless of the objective.

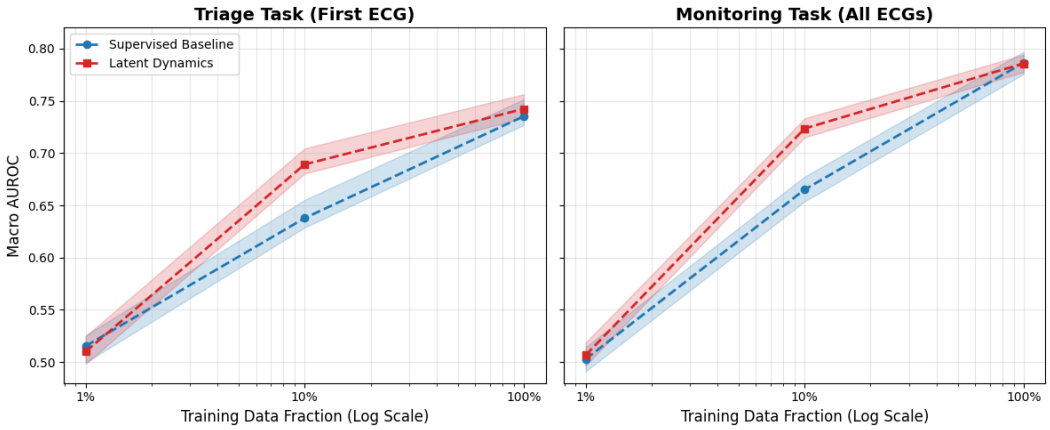


Figure 4: Macro-AUROC vs. Data Fraction. Results are shown for Triage (First ECG) and Monitoring (All ECGs) tasks. The Latent Dynamics model demonstrates significantly higher sample efficiency in the 10% regime ($p < 0.05$, non-overlapping intervals) compared to the fully supervised baseline.

**Failure-mode.** To investigate this performance degradation at the 1% data regime, we conduct a post-mortem analysis comparing full-parameter finetuning against linear probing on frozen representations. Results are detailed in Table 2. We observe a striking divergence: while full-parameter finetuning collapses to random guessing, the linear probe maintains a significantly higher signal.

This suggests that the pretraining successfully captures cardiac dynamics and organizes a meaningful latent space even at this extreme low-data scale. However, the standard supervised finetuning objective appears to introduce sufficient gradient noise to overwrite these learned features when labels are sparse. This finding confirms that the collapse is not a failure of the world model's representation learning, but rather a failure of the task-specific adaptation protocol, highlighting the need for more robust distillation methods.

## 5 Conclusion and Future Work

We presented an Action-Conditioned World Model for cardiac monitoring that learns to simulate the electrophysiological evolution of the heart. By decoupling patient identity from pathological transitions via the LeJEPA architecture, our method achieves parity with fully supervised baselines on longitudinal monitoring while slightly outperforming them on the critical Triage task (0.742 vs 0.735 AUROC). Most notably, in low-resource regimes (10% data), our dynamics objective serves as a potent regularizer, improving performance by over 0.05 points with confidence, showing evidence

Table 2: Low-Data Regime Analysis (1% Data). We compare the Supervised baseline against our World Model using two evaluation protocols. The gap between Linear Probe and Finetuned results indicates that the world model retains clinical features that are destroyed during full-parameter finetuning.

| | | Macro-AUROC | |
|---|---|---|---|
| **Method** | **Eval. Protocol** | **Triage (First ECG)** | **Monitoring (All)** |
| Supervised | End-to-End | 0.515 [0.500 – 0.526] | 0.503 [0.491 – 0.514] |
| Ours (Latent Dynamics) | Linear Probe | 0.563 [0.551 – 0.575] | 0.575 [0.560 – 0.591] |
| | Finetuned | 0.510 [0.498 – 0.525] | 0.507 [0.497 – 0.519] |

that the dense signal of predicting future states, given the label-action, is more sample-efficient than static classification. For future work, we envision three key directions to scale this framework:

**Data efficiency and the limits of label-as-action.** Since our predictor approximates a reversible dynamical system, we can generate synthetic training data by reversing transitions ($h_{t+1}, -a_t \to \hat{h}_t$) or chaining multiple pathology vectors, potentially unlocking performance in the 1% data regime where current models collapse. However, the validity of reversal is not uniform across pathologies and opens a deeper question about the representability of $a_t$ itself. Disease progression is fundamentally entropic and temporally directed (Agrawal et al., 2022). For example, the resolution of a paroxysmal arrhythmia (I48) is a plausible clinical trajectory and its reversal, onset from a clean baseline, is a symmetric and informative counterfactual. By contrast, reversal is physically inadmissible for irreversible conditions such as myocardial infarction (I21), where the *arrow of time* is strictly one-directional. Crucially, this asymmetry reveals a broader limitation of the binary action space $\{-1, 0, 1\}^C$, it encodes only the *fact* of a transition, not its *clinical magnitude*. An acute arrhythmia onset and a gradual chronic drift are aliased to the same $a_{t,i} = 1$ regardless of their fundamentally different dynamics.

This aliasing is most consequential for the chronic conditions that dominate our dataset. Conditions such as Hypertension (I10) and Ischemic Heart Disease (I25), the two most frequent labels in MIMIC-IV-ECG, progress over years rather than timesteps, meaning the transition vector will be near-zero for the vast majority of their sequential pairs. The world model, therefore, might be well-regularized for discrete acute events but receives a weak and infrequent training signal for the slow drift of chronic substrates. Furthermore, in real clinical data the precise onset of even acute pathologies is often temporally ambiguous. ICD-10 codes may be assigned retrospectively, and the physiological change captured in the waveform may precede or lag the recorded label. We note that MIMIC-IV-ECG mitigates this by pairing each waveform with codes assigned at the time of acquisition, and that the prevalence of sparse single-label transitions (refer to Appendix A) is consistent with discrete clinical events rather than noisy multi-label drift, but the ambiguity cannot be fully eliminated within a binary action space.

Both limitations might converge to a single resolution. Future iterations can replace the one-hot transition vector with a learned continuous Action Embedding ($a_t \in \mathbb{R}^d$). A dense embedding space can represent gradations of onset, encode temporal asymmetry by construction, and assign geometrically proximal vectors to clinically related conditions, allowing rare or slowly-evolving diseases to borrow statistical strength from common ones. Conditioning the dynamics predictor additionally on the *elapsed time* $\Delta t$ between recordings would further disentangle the two regimes: a one-day gap where stability is expected versus a one-year gap where chronic drift is plausible. This extension aligns naturally with Neural ODE formulations (Wen & Biros, 2025), where the latent trajectory is a continuous function of time, enabling the same framework to capture both the sharp electrophysiological signatures of acute events and the slow, irreversible drift of chronic disease substrates without requiring separate model families for each regime.

**Label sparsity.** We currently restrict labels to Chapter IX (Circulatory) truncated to 2 digits. The learned Action Embedding described above scales this naturally to the full ICD-10 ontology: unlike supervised heads which require dense samples per class to converge, a dynamics model shares

the transformation physics across the embedding space, allowing rare diseases to borrow statistical strength from common ones without the instability typical of high-dimensional classification heads.

**Counterfactual Simulation.** Leveraging the geometry enforced by SIGReg, we aim to deploy the model as a patient-specific simulator. By applying a disease vector to a healthy patient's embedding ($\hat{h}_{sick} = \text{Dyn}_\phi(h_{healthy}, a_{disease})$), we can generate counterfactual trajectories—answering *"What would this specific patient look like if they developed Atrial Fibrillation?"*—offering a powerful tool for interpretability. ¡- contextualized inference

## REFERENCES


Monica N Agrawal, Hunter Lang, Michael Offin, Lior Gazit, and David Sontag. Leveraging time irreversibility with order-contrastive pre-training. In *International Conference on Artificial Intelligence and Statistics*, pp. 2330–2353. PMLR, 2022.

Matilda Andersson. Self-supervised representation learning from electrocardiogram data for medical applications. *Master's Theses in Mathematical Sciences*, 2023.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3478–3488, 2021.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D Aguirre, Collin M Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLoS computational biology*, 18(2):e1009862, 2022.

Mouad El Bouchattaoui, Myriam Tami, Benoit Lepetit, and Paul-Henry Cournède. Causal contrastive learning for counterfactual regression over time. *Advances in Neural Information Processing Systems*, 37:1333–1369, 2024.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, et al. Mimic-iv-ecg: Diagnostic electrocardiogram matched subset. *Type: dataset*, 6:13–14, 2023.

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

Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Computers in biology and medicine*, 141:105114, 2022.

Manuel T Nonnenmacher, Lukas Oldenburg, Ingo Steinwart, and David Reeb. Utilizing expert features for contrastive learning of time-series representations. In *International Conference on Machine Learning*, pp. 16969–16989. PMLR, 2022.

Chaoxu Ren, Le Sun, and Dandan Peng. A contrastive predictive coding-based classification framework for healthcare sensor data. *Journal of Healthcare Engineering*, 2022(1):5649253, 2022.

Shiva Salsabilian and Laleh Najafizadeh. Subject-invariant feature learning for mtbi identification using lstm-based variational autoencoder with adversarial regularization. *Frontiers in Signal Processing*, 2:1019253, 2022.

Nisarg A Shah, Mingze Xia, Shameema Sikder, S Swaroop Vedula, and Vishal M Patel. Learning action-conditioned world models for cataract surgery from unlabeled videos. In *Medical Imaging with Deep Learning*, 2026.

Nils Strodthoff, JM Lopez Alcaraz, and W Haverkamp IV. Mimic-iv-ecg-ext-icd: Diagnostic labels for mimic-iv-ecg (version 1.0. 1). *PhysioNet. RRID: SCR_007345 https://doi. org/10.13026/hdyc-1h77*, 2024a.

Nils Strodthoff, Juan Miguel Lopez Alcaraz, and Wilhelm Haverkamp. Prospects for artificial intelligence-enhanced electrocardiogram as a unified screening tool for cardiac and non-cardiac conditions: an explorative study in emergency care. *European Heart Journal-Digital Health*, 5 (4):454–460, 2024b.

Zheyu Wen and George Biros. Lnode: Uncovering the latent dynamics of a $\beta$ in alzheimer's disease. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 313–322. Springer, 2025.

Erik Xie, Wyatt Chang, Raquel Rodriguez Martinez, and Brandon Ballinger. Jets: A self-supervised joint embedding time series foundation model for behavioral data in healthcare. In *NeurIPS 2025 Workshop on Learning from Time Series for Health*, 2025.

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.

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 Computer Vision and Pattern Recognition Conference*, pp. 20778–20788, 2025.

Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, et al. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE transactions on pattern analysis and machine intelligence*, 46(10):6775–6794, 2024.

Qingyu Zhao, Zixuan Liu, Ehsan Adeli, and Kilian M Pohl. Longitudinal self-supervised learning. *Medical image analysis*, 71:102051, 2021.

Ya Zhou, Yujie Yang, Jianhuang Gan, Xiangjie Li, Jing Yuan, and Wei Zhao. Multi-scale masked autoencoder for electrocardiogram anomaly detection. *arXiv preprint arXiv:2502.05494*, 2025.

## A Dataset Characterization

We evaluate our model using the MIMIC-IV-ECG diagnostic matched subset, a large-scale longitudinal repository of clinical 12-lead ECG waveforms. This appendix provides a detailed statistical characterization of the dataset, the label processing workflow, and the resulting distribution of pathological transitions (actions).

**Data Scale and Longitudinality.** The dataset contains a total of $800,035$ ECG records associated with $161,352$ unique patients. A defining feature of this dataset is its longitudinal nature, which is essential for training world models. While $57,344$ patients are represented by only a single recording (singletons), the majority of the cohort ($104,008$ patients) have two or more recordings. On average, each patient has approximately $4.96$ ECGs (median: $2.0$; maximum: $260$). See Figure 5.

By extracting consecutive recordings for each patient, we identified $638,683$ unique sequential pairs $(X_t, X_{t+1})$. These pairs serve as the primary unit of observation for our action-conditioned pretraining, where the model learns the transition dynamics from $t$ to $t+1$.

**Label Processing and Cardiac State Definition.** Each ECG in the dataset is annotated with ICD-10 codes, with an average of $8.15$ codes per record across $15,197$ unique medical conditions. To focus the world model on cardiac electrophysiology, we filtered these labels for ICD-10 Chapter IX (Diseases of the Circulatory System), identified by the prefix 'I'.

Following standard clinical grouping, we truncated these codes to their first three digits (e.g., grouping 'I48.91' and 'I48.1' into the broader category 'I48' for Atrial Fibrillation/Flutter). This reduction

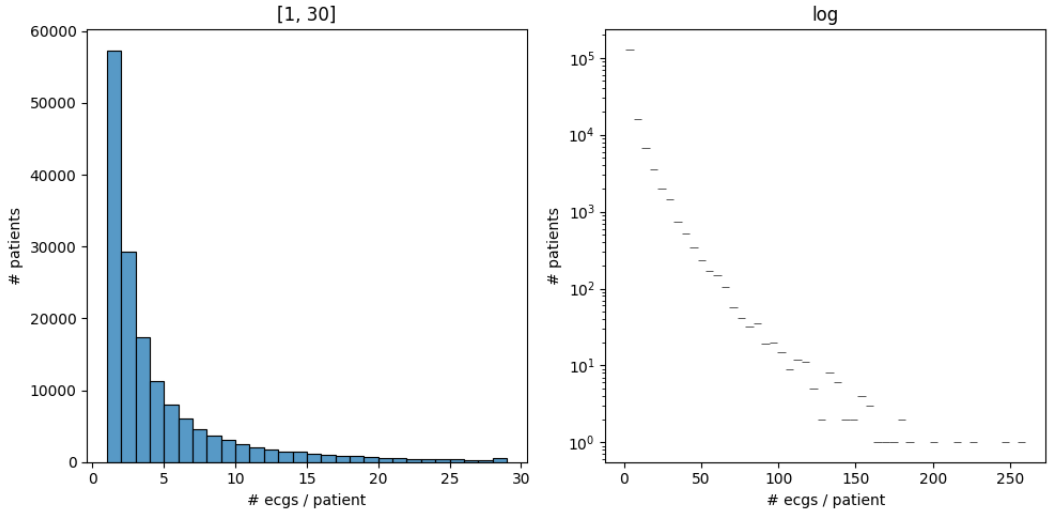


Figure 5: Distribution of longitudinal ECG recordings per patient. The left panel shows a histogram for the bulk of the cohort (1 to 30 recordings per patient), demonstrating that while many patients are singletons, the majority contribute multiple observations. The right panel provides a full-scale view on a logarithmic y-axis, highlighting the dataset's significant long tail.

resulted in an action space of 76 distinct cardiac clusters. Under this specialized label set, the average number of cardiac-specific codes per ECG is $1.45$. The most frequent clusters in the dataset include I10 (Hypertension), I25 (Ischemic Heart Disease), I48 (Atrial Fibrillation/Flutter), and I50 (Heart Failure).

**Characterizing Pathological Actions.** We define an "Action" as the transition vector $\boldsymbol{a}_t = \boldsymbol{y}_{t+1} - \boldsymbol{y}_t$. Across the $638,683$ sequential pairs, we observed the following distribution of dynamics, also see Figure 6:

- **Stable Pairs (Identity):** $465,842$ pairs ($72.9\%$) showed no change in cardiac ICD-10 clusters between time steps. These pairs provide the "identity" signal, forcing the encoder to learn representations that remain invariant when no action occurs.
- **Changed Pairs (Action):** $172,841$ pairs ($27.1\%$) involve a change in the cardiac state. These transitions are contributed by $55,808$ unique patients, with a mean of $3.10$ "changed" pairs per contributing patient.

The "magnitude" of these actions (defined as the $L_1$ norm of the transition vector) is typically sparse. Among the changed pairs, the most common magnitude is a single label transition (adding or removing one disease state), which accounts for $32.6\%$ of all actions. Changes involving 2 or 3 labels account for $22.8\%$ and $18.1\%$ respectively. This sparsity supports our formulation of disease onset as a translational force within the latent space, where most clinical events represent specific, localized movements on the latent manifold.

## B Experimental Setup

**Hardware and Environment.** All experiments were conducted on a single node equipped with an NVIDIA RTX 4090 (24GB VRAM). To mitigate the I/O bottlenecks inherent to processing $\sim$800k small waveform files, we consolidated the dataset into a contiguous HDF5 monolith (stored in Float16 precision) and utilized a persistent RAM caching strategy for the low-data regimes.

**Architecture and Optimization.** We utilized an xResNet1d50 backbone. For the LeJEPA pretraining, the projector was a 3-layer MLP expanding the 256-dim backbone output to a 256-dim

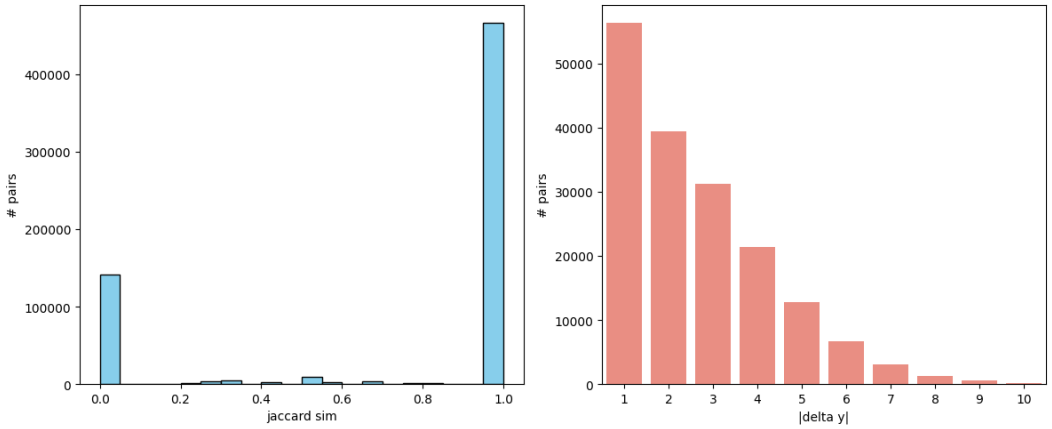


Figure 6: Characterization of cardiac label transitions. The left panel shows the distribution of Jaccard similarity between consecutive label sets $(y_t, y_{t+1})$. The prominent peak at 1.0 represents stable pairs with identical cardiac states, while the remainder identifies the changed pairs in the dataset. The right panel details the magnitude of these actions ($L_1$ norm of the transition vector) for all changed pairs. The prevalence of sparse transitions—specifically single-label changes ($32.6\%$)—supports the formulation of disease onset as a localized translational force in the latent manifold.

Table 3: Regularization Ablation (Full Dataset). Comparison of the Latent Dynamics model pretrained with VICReg versus SIGReg. Results are reported for the fine-tuned evaluation protocol. SIGReg outperforms VICReg on both tasks, validating the choice of Isotropic Gaussian geometry over the hypercube constraints of VICReg for physiological signals.

| | **Macro-AUROC** | |
|---|---|---|
| **Regularization** | **Triage (First ECG)** | **Monitoring (All)** |
| VICReg | 0.720 [0.710 – 0.738] | 0.770 [0.759 – 0.780] |
| SIGReg | 0.742 [0.733 – 0.756] | 0.786 [0.777 – 0.794] |

latent space (matching the backbone capacity to avoid information bottlenecks). The predictor was a residual MLP with hidden dimension 512. Training utilized the AdamW optimizer with a OneCycleLR scheduler. We observed that aggressive learning rates (max LR $1e^{-3}$) were stable for the dynamics task, whereas the supervised baseline required lower rates ($1e^{-4}$) and gradient clipping to prevent divergence. The batch size was set to 512 to maximize GPU throughput, which necessitated careful scaling of the loss terms.

**Regularization Ablation.** Prior to adopting SIGReg, we attempted to train the dynamics model using the VICReg objective. As shown in Table 3, this approach yielded suboptimal performance compared to SIGReg. We attribute this to VICReg's sensitivity to its three hyperparameters $(\lambda, \mu, \nu)$, which typically require extensive grid search to balance. Due to computational constraints, we adopted a heuristic tuning strategy: monitoring the gradient magnitude ratio between the Prediction and Regularization terms during the first epoch. While this allowed us to stabilize SIGReg by simply scaling its single parameter $\lambda$, achieving a similar equilibrium with the multi-term VICReg objective proved intractable on this modality.