CARDIACSTATE: LEARNING CYCLIC CARDIAC STATE THROUGH PREDICTIVE VIDEO PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting how an observed physical system will evolve requires inferring the underlying state that governs its motion. While controllable video generation relies on external action or text signals, action-free predictive models leave this conditioning implicit within high-dimensional representations. Autonomous biological systems, such as the beating heart, lack external controllers—their driving dynamics unfold continuously within the visual stream itself. Yet, core clinical tasks like keyframe localization and Left Ventricular Ejection Fraction (LVEF) estimation depend heavily on tracking this physiological state. We introduce CardiacState, a joint-embedding predictive architecture (JEPA) that addresses this by formulating physiological progression as an explicit, self-derived conditioning state on a continuous cyclic manifold S^1. By imposing a causal temporal bottleneck that denies the future-feature predictor direct access to historical frames, CardiacState forces all temporal context to route exclusively through this S^1 state alongside physical elapsed time $\Delta\tau_t$. Pretrained on 1.88 million clinical echocardiogram videos without external text, action labels, or hardware ECG gating, CardiacState advances both feature representation and structural interpretability. Its frozen visual representations achieve $4.87 \pm 0.12\%$ LVEF MAE on EchoNet-Dynamic and $4.73 \pm 0.18\%$ on EchoNet-Pediatric—outperforming prior self-supervised baselines (5.76% and 5.12%). Furthermore, unlike implicit representations, this explicit cyclic state directly enables non-parametric keyframe localization across diverse acoustic views via phase-anchor matching, reducing mean localization error from 2.25 frames to 1.99 ± 0.02 frames. Controlled ablations show that cardiac progression regularization is important for learning a useful cyclic coordinate, while the temporal bottleneck provides further gains in localization and representation quality. These results support cyclic state conditioning for predictive learning in quasi-periodic systems.

1 INTRODUCTION

Echocardiography serves as the primary non-invasive modality for evaluating cardiac function, which clinically relies on tracking cyclic left ventricular (LV) deformation across complete heartbeats—most notably quantified by Left Ventricular Ejection Fraction (LVEF) via End-Diastolic (ED) and End-Systolic (ES) keyframes (Lang et al., 2015; Ouyang et al., 2020). However, isolated ultrasound frames reveal only static anatomical snapshots; virtually identical ventricular geometries recur during both contraction and relaxation (Chen et al., 2026). Because visual appearance alone cannot determine the phase or direction of cardiac motion, standard visual encoders struggle to distinguish structurally identical but functionally distinct frames. This motivates a central question in medical video representation learning: **can a model infer the underlying physiological state explicitly from visual dynamics alone, without relying on external labels?**

Conditioned prediction offers a compelling paradigm for modeling temporal dynamics (Ha & Schmidhuber, 2018; Zhou et al., 2024). In robotics and controllable video synthesis, future prediction relies on externally supplied actions (Zhu et al., 2026) or textual prompts (Wu et al., 2026). Conversely, action-free predictive architectures—such as V-JEPA (Bardes et al., 2024; Assran et al., 2025)—forecast future features without external controls, but leave the driving conditioning signal *implicit* within high-dimensional visual context. Autonomous biological processes represent a distinct regime: the beating heart lacks external operators, yet its driving condition is an explicit, low-dimensional

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.

physical state that unfolds continuously within the visual stream. While latent-action world models infer unobserved controls from video (Schmidt & Jiang, 2024; Bruce et al., 2024), they formulate these signals as abstract, unconstrained codes. We instead derive a physically grounded conditioning state from video, representing cardiac progression on S^1.

Self-supervised video pretraining provides a natural framework for learning representations from large, unannotated clinical cine loops (Böhi et al., 2026; **?**). Pixel-reconstruction models like VideoMAE (Tong et al., 2022) remain sensitive to ultrasound acoustic speckle, whereas Joint-Embedding Predictive Architectures (JEPAs) circumvent pixel decoding by predicting target features in latent space (Munim et al., 2026). However, robust feature prediction does not automatically yield an explicit coordinate for cardiac progression. Because echocardiographic sequences exhibit strong short-range visual continuity, standard predictive encoders can exploit local persistence shortcuts—predicting future features by extrapolating nearby pixel movement while leaving learned representations weakly organized around global cardiac mechanics.

Overcoming these shortcuts requires incorporating three physical constraints intrinsic to echocardiographic motion: (1) **Visual Phase Ambiguity:** resolving directional contraction vs. relaxation; (2) **Topological Recurrence:** mapping smooth periodic trajectories onto a circular manifold S^1 (Laumer et al., 2020); and (3) **Acquisition Heterogeneity:** accommodating patient heart rate variations and frame-rate differences via physical time elapsed $\Delta\tau_t$. To address these challenges, we introduce CardiacState, a JEPA framework that explicitly parameterizes visual history on S^1 (Figure 1). Crucially, CardiacState enforces a strict temporal bottleneck: the future-feature predictor is granted direct access only to the single current frame feature ($K = 1$), forcing all historical context to route exclusively through the S^1 cardiac state and physical elapsed time $\Delta\tau_t$. By denying the predictor direct access to historical visual tokens, CardiacState supresses local persistence shortcuts and forces the learned state to encode global physiological progression.

Our primary contributions are summarized as follows:

1. **Explicit Cardiac State Formulation for Video JEPAs:** We introduce CardiacState, a self-supervised video predictive framework that parameterizes continuous cardiac state on a 1D circular manifold S^1. In contrast to latent-action world models with abstract control codes and action-conditioned generators, CardiacState derives an internal, physically interpretable conditioning state directly from visual history. Cardiac progression regularization organizes this state into a forward-moving cyclic coordinate; a causal temporal bottleneck ($K = 1$) then makes it the predictor's route to visual history. Controlled ablations show substantial localization gains from the progression loss and further localization and representation gains from the bottleneck.
2. **ECG-Free Keyframe Localization from Frozen States:** We demonstrate that the learned S^1 state space enables non-parametric localization of End-Diastolic (ED) and End-Systolic (ES) keyframes. Using frozen states and calibrated phase anchors, CardiacState reduces mean ED/ES localization error from 2.245 to 2.002 frames across three views.
3. **Enhanced Visual Representations for Functional Assessment:** We show that forcing visual features to align with global cardiac progression improves downstream feature utility. Pretrained on 1.88 million clinical videos, CardiacState achieves the lowest LVEF estimation MAE among visual encoders evaluated under our frozen-backbone attention-probing protocol, reducing MAE from 5.19% to 4.87% on EchoNet-Dynamic and from 4.81% to 4.73% on EchoNet-Pediatric.

2 Related Work

Predictive video representations. VideoMAE reconstructs masked video pixels (Tong et al., 2022), and spatiotemporal contrastive learning aligns representations across augmented clips (Qian et al., 2021). Joint-embedding predictive architectures instead predict target features from observed context (Bardes et al., 2024; Assran et al., 2025). V-JEPA 2 pretrains without actions and subsequently adapts a predictor for action-conditioned robotic planning (Assran et al., 2025). CardiacState also predicts future features, but infers a cyclic conditioning state during pretraining and makes that state the predictor's only access to visual history beyond the current frame.

Disambiguating futures through conditioning. Similar appearances can precede different futures, so prediction depends on what information identifies the current dynamics. Astra conditions video

generation on supplied actions (Zhu et al., 2026); Aurora uses optional text and image context to guide time-series forecasting (Wu et al., 2026). More closely related, imitation from observation and latent-action models infer transition-relevant variables without observed action labels (Edwards et al., 2019; Schmidt & Jiang, 2024; Bruce et al., 2024). Their learned variables support policy learning or controllable generation. In echocardiography, similar ventricular appearances can occur during contraction and relaxation. CardiacState infers an endogenous cyclic state from preceding motion and uses it with elapsed time to condition future-feature prediction. Its S^1 constraint and cardiac progression objective make this state a phase coordinate, which we evaluate directly through keyframe localization.

Echocardiographic representation learning and acoustic robustness. Supervised video models established automated LVEF estimation (Ouyang et al., 2020), and PanEcho extends clinical prediction across multiple tasks (Holste et al., 2025). EchoCLIP and EchoPrime connect echocardiographic images or videos with clinical reports (Christensen et al., 2024; Vukadinovic et al., 2026). Among video-pretraining methods, Echo-Vision-FM uses masked autoencoding at scale followed by task-specific fine-tuning (Zhang et al., 2025); its reported LVEF results use a different protocol from our frozen-encoder probe. Video-LAMAE and EchoCardMAE adapt masking and reconstruction to cardiac video (Böhi et al., 2026; **?**). EchoJEPA predicts latent features and reports robustness to simulated attenuation and shadowing (Munim et al., 2026), while US-JEPA studies speckle corruption in broader ultrasound imaging (Radhachandran et al., 2026). CardiacState focuses on organizing cardiac progression as an explicit conditioning state while learning future features from videos and acquisition timing, starting from a DINOv3 visual encoder (Siméoni et al., 2025).

Cyclic state learning and cardiac keyframe localization. Cardiac phase and ED/ES keyframes have been estimated with supervised CNNs, recurrent models, optical flow, and hierarchical motion models (Farhad et al., 2023; Lane et al., 2021; Lu et al., 2025; Chen et al., 2026). Self-supervised alternatives recover cardiac motion trajectories through reconstruction (Laumer et al., 2020; Yang et al., 2025); recent single-parameter latent orbits impose an explicit phase prior for keyframe detection (Bonnici et al., 2026). CardiacState likewise uses a cyclic prior, but trains the state to condition future-feature prediction under a temporal bottleneck, linking phase learning to visual representation pretraining. After training, ED/ES localization uses the frozen state and phase anchors calibrated from training annotations, without a task-specific detection head or ECG gating.

3 CardiacState: Predictive Representation Learning via Cyclic State

To enable conditioned prediction without external control signals, CardiacState infers an internal physiological state directly from visual dynamics to serve as a self-derived conditioning signal. Figure 1 illustrates how image encoding (a), history-dependent cardiac phase prediction (b), and future-feature prediction (c) realize this framework. Rather than allowing the predictor direct access to multi-frame visual history, CardiacState routes historical context exclusively through an explicit state pathway on the unit circle S^1. The predictor combines instantaneous visual structure with this inferred cyclic cardiac state c_t and physical elapsed time $\Delta\tau_t$ to forecast future latent features. After pretraining, cosine comparison to reference phase anchors transforms the learned latent state into a keyframe localization readout (d).

3.1 A History-Dependent Coordinate of Cardiac Progression

Let $x_{1:T}$ denote an echocardiographic cine clip acquired at times $\tau_{1:T}$, where $\Delta\tau_t = \tau_{t+1} - \tau_t$ represents physical elapsed time in milliseconds. An image encoder E_ϕ and projection head q_ξ map each frame to an image feature vector $F_t = q_\xi(E_\phi(x_t)) \in \mathbb{R}^d$. We initialize the visual encoder from DINOv3 (Siméoni et al., 2025) and project its normalized class token into the predictive feature space. Architectural dimensions are provided in Appendix A.

Resolving phase ambiguity on S^1. Single frame suffers from *Visual Phase Ambiguity*: myocardial geometry during mid-systole and mid-diastole often appears structurally identical, making directional trajectory impossible to determine from an isolated frame. Furthermore, representing phase as a scalar

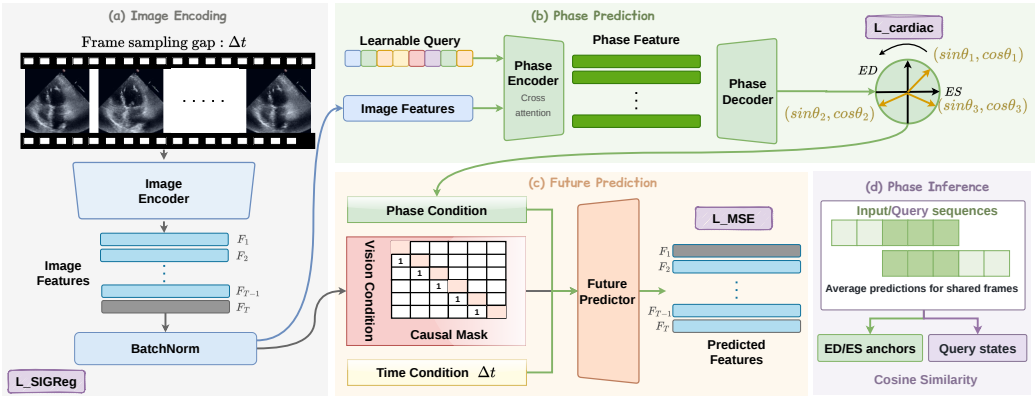


Figure 1: **CardiacState architecture and downstream readouts.** (a) Image Encoding: A shared image encoder extracts frame features F_t, regularized by $\mathcal{L}_{\text{SIGReg}}$ after projection and Batch Normalization. (b) Phase Prediction: Learnable queries summarize historical visual context through cross-attention into an explicit cardiac state $c_t = (\sin\theta_t, \cos\theta_t) \in S^1$. (c) Future Prediction: Conditioned on cardiac state c_t and physical elapsed time $\Delta\tau_t$, the future predictor receives only the current visual feature under a strict causal bottleneck mask M_1 to predict $\widehat{F}_{t+1}$, optimized via $\mathcal{L}_{\text{MSE}}$. The sampling gap labeled Δt in the diagram denotes the elapsed physical time $\Delta\tau_t$ used in the equations. (d) Phase Inference: Overlapping sliding windows yield repeated state estimates for shared frames, whose cosine scores against training-calibrated ED/ES phase anchors are averaged to localize keyframes without hardware ECG gating.

on a bounded interval introduces a wrap-around boundary at each cycle transition. Parameterizing the latent cardiac state vector c_t on the 1D unit circle S^1 addresses this *Topological Recurrence*.

To aggregate historical context, the cardiac phase module g_ψ (Figure 1b) processes visual features $F_{1:t}$ through a causal pathway. The cardiac phase encoder uses causal cross-attention where learnable queries aggregate historical representations. The output features pass through a cardiac phase decoder to produce $r_t \in \mathbb{R}^2$, which is unit-normalized onto S^1 to yield the explicit 2D cardiac state vector c_t:

$$c_t = g_\psi(F_{1:t}) = \frac{r_t}{\|r_t\|_2} = \begin{bmatrix} \sin\theta_t \\ \cos\theta_t \end{bmatrix} \in S^1, \tag{1}$$

where $\theta_t = \text{atan2}(c_{t,1}, c_{t,2})$. Aggregating historical visual context $F_{1:t}$ enables the cardiac state c_t to resolve visual phase ambiguities, while the circular topology smoothly connects the end of one cardiac cycle to the start of the next.

Enforcing forward temporal progression. Unsupervised state estimation risks representation collapse, where the model outputs stationary states or oscillates backward. To enforce physical temporal ordering while permitting biological velocity variations (e.g., patient-specific heart rates), we compute wrapped angular increments:

$$\delta\theta_t = \text{atan2}(\sin(\theta_{t+1} - \theta_t),\ \cos(\theta_{t+1} - \theta_t)), \tag{2}$$

and penalize angular progression that falls below a time-scaled physical threshold ϵ:

$$\mathcal{L}_{\text{cardiac}} = \frac{1}{T-2} \sum_{t=1}^{T-2} [\epsilon\Delta\tau_t - \delta\theta_t]_+, \qquad \epsilon = 10^{-3}\ \text{rad/ms}. \tag{3}$$

For $T-1$ state predictions, this hinge loss evaluates $T-2$ adjacent state transitions. We choose $\epsilon = 10^{-3}\,\text{rad/ms}$ as a permissive progression margin. For scale, a nominal 60-bpm beat lasts 1000 ms, so one turn over that interval would average $2\pi/1000 \approx 6.3 \times 10^{-3}\,\text{rad/ms}$, about six times our margin. The loss thus penalizes nearly stationary or backward increments without prescribing a constant angular speed; this flexibility is motivated by the heart-rate-dependent durations of cardiac subphases (Chung et al., 2004). The calculation sets a scale for the regularizer, not a requirement that the learned state complete one turn per beat.

3.2 A Temporal Bottleneck for Future-Feature Prediction

Routing historical context through the cardiac state. Echocardiographic sequences exhibit high frame-to-frame visual persistence; adjacent frames share nearly identical myocardial geometry. Predictors given multi-frame visual history can use this short-range continuity to forecast future features without explicitly representing global cardiac progression.

During pretraining, we sample 16-frame clips at a stride $s \in \{2, 3, 4, 5, 6\}$ in original-video frames. At acquisition frame rate f, sampled observations are $\Delta\tau_t = 1000s/f$ ms apart. Varying s and f weakens local-persistence shortcuts while giving the predictor elapsed physical time.

To couple the inferred cardiac state to future-feature prediction, we introduce a condition fuser network $h_\eta(c_t, \Delta\tau_t)$ that combines the history-derived cardiac state c_t with physical elapsed time $\Delta\tau_t$. We then impose a strict temporal bottleneck mask M_1 on the future predictor P_ω:

$$\widehat{F}_{t+1} = P_\omega(F_t,\ h_\eta(c_t, \Delta\tau_t)) \quad \text{subject to causal mask } M_1. \tag{4}$$

The mask M_1 restricts visual attention solely to the single current frame feature F_t (Figure 1c). Consequently, the predictor retains direct access to instantaneous anatomical structure, but is strictly denied direct visual history. To predict the target feature $\widehat{F}_{t+1}$, the network is forced to route all historical context through $h_\eta(c_t, \Delta\tau_t)$.

Handling heterogeneity via condition fuser h_η. To address *Temporal and Acquisition Heterogeneity* (where acquisition frame rates and patient heart rates vary), h_η explicitly incorporates physical elapsed time $\Delta\tau_t$ alongside the cardiac state c_t. The cardiac state vector $c_t \in \mathbb{R}^2$ is mapped through an MLP, while $\Delta\tau_t$ in milliseconds is mapped via sinusoidal temporal embeddings followed by an MLP. Learned scalar weights combine both branches before Layer Normalization. Following LeWM (Maes et al., 2026), the fused condition $h_\eta(c_t, \Delta\tau_t)$ is injected via Adaptive Layer Normalization (AdaLN) (Peebles & Xie, 2023; Maes et al., 2026) as scale, shift, and gate parameters. Fusing physical elapsed time in milliseconds ensures the model predicts feature evolution across true physical time intervals rather than arbitrary frame indices.

3.3 Joint Learning from Video and Acquisition Timing

For B clips, future-feature prediction minimizes latent-space mean squared error:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{B(T-1)d} \sum_{b=1}^{B} \sum_{t=1}^{T-1} \|\widehat{F}_{b,t+1} - F_{b,t+1}\|_2^2. \tag{5}$$

The total self-supervised pretraining objective is:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + 0.1\mathcal{L}_{\text{SIGReg}} + \mathcal{L}_{\text{cardiac}}, \tag{6}$$

where $\mathcal{L}_{\text{SIGReg}}$ denotes the SIGReg regularization loss introduced by Balestriero & LeCun (2025) and adopted in LeWM (Maes et al., 2026). It matches the empirical characteristic function of projected latent features to a standard multivariate Gaussian over the global batch. This term prevents representation collapse and preserves feature variance across spatial and temporal dimensions without negative pairs, memory banks, or pixel-level decoding. Its weight follows Maes et al. (2026).

Crucially, the entire pretraining objective operates in a purely self-supervised manner, relying exclusively on raw video frames and acquisition frame rates—without requiring hardware ECG signals, clinical phase annotations, or dense anatomical masks.

3.4 Two Complementary Readouts of the Pretrained Representation

To evaluate the utility of the learned representations, we examine two distinct readouts without fine-tuning the underlying backbone.

Functional assessment through frozen visual features. To verify that feature prediction forces the visual encoder E_ϕ to capture global functional dynamics, we strictly freeze E_ϕ and evaluate

downstream Left Ventricular Ejection Fraction (LVEF) regression. A trainable attention probe aggregates patch tokens from uniformly sampled video frames using spatial and temporal position encodings, then regresses LVEF from a global token. Appendix A specifies the readout architecture. The cardiac phase module, future predictor, and projection heads are detached; downstream accuracy evaluates the physiological representation learned by E_ϕ itself.

Non-parametric keyframe localization from frozen cyclic states. Existing approaches for cardiac keyframe identification predominantly rely on task-specific supervised detectors (Chen et al., 2026) or explicit motion modeling (Farhad et al., 2023; Yang et al., 2025), frequently requiring hardware ECG gating for temporal alignment. In contrast, CardiacState utilizes its frozen cyclic state space $\mathbf{c}_t \in S^1 \subset \mathbb{R}^2$ as a direct, interpretable coordinate for keyframe localization. By matching frame-level state representations to dataset-calibrated phase anchors via vector dot products, temporal keyframe detection becomes a simple non-parametric readout of the frozen manifold—performing precise phase tracking without hardware ECG signals, external control inputs, or trained task heads.

Specifically, an unannotated query video is processed using overlapping 16-frame sliding windows (stride 2, 50% overlap) to assign each frame t a unit state vector $\mathbf{c}_t \in \mathbb{R}^2$. From training annotations, we calibrate two unit reference phase anchors on S^1 via circular vector averaging: End-Diastole ($\mathbf{a}_{\text{ED}}$) and End-Systole ($\mathbf{a}_{\text{ES}}$). For a test clip, frame-level phase alignment scores are obtained:

$$S_{\text{ED}}(t) = \mathbf{c}_t^\top \mathbf{a}_{\text{ED}}, \quad S_{\text{ES}}(t) = \mathbf{c}_t^\top \mathbf{a}_{\text{ES}}. \tag{7}$$

Because $\mathbf{c}_t$ and $\mathbf{a}_k$ are unit-normalized, $S_k(t) \in [-1, 1]$ directly measures angular proximity to the respective cardiac phase anchor. End-Diastolic and End-Systolic keyframes are then localized by extracting the local maxima of $S_{\text{ED}}(t)$ and $S_{\text{ES}}(t)$, respectively, within a 5-frame local temporal window (Chen et al., 2026). Details are provided in Appendix C.

4 Experimental Evaluation

Our experiments evaluate whether routing temporal history exclusively through an explicit cyclic state forces the visual encoder to capture global physiological dynamics, and whether the inferred state vector c_t encodes cardiac progression.

4.1 Experimental Setup

Pretraining corpus & compute setup. We pretrain CardiacState on a private dataset of 1,883,012 clinical echocardiography videos possessing valid frame-rate metadata and containing at least 32 frames. This private pretraining dataset does not overlap with the EchoNet-Dynamic, EchoNet-Pediatric, or EchoNet-LVH datasets. Each pretraining clip consists of 16 frames sampled uniformly at a stride selected from $\{2, 3, 4, 5, 6\}$. Multiplying the frame stride by recorded acquisition timing yields the true physical interval $\Delta\tau_t$ in milliseconds. Pretraining runs for 10 epochs with a global batch size of 128 on 8 $\times$ NVIDIA H200 GPUs, taking approximately 32 wall-clock hours.

Evaluation benchmarks. We evaluate downstream functional assessment and keyframe localization across three established echocardiography benchmarks: **EchoNet-Dynamic** (Ouyang et al., 2020): Contains 10,030 adult Apical 4-Chamber (A4C) videos with expert LVEF labels and ES/ED keyframe annotations. **EchoNet-Pediatric** (Reddy et al., 2023): Contains 7,643 pediatric A4C and Parasternal Short-Axis (PSAX) videos with expert LVEF annotations. **EchoNet-LVH** (Duffy et al., 2022): Contains 12,000 Parasternal Long-Axis (PLAX) videos with clinical keyframe labels. We strictly maintain official train/validation/test splits across all benchmarks.

Baselines & evaluation protocols. For LVEF regression, we compare against general-purpose visual encoders (DINOv3 (Siméoni et al., 2025)), task/report-supervised echocardiography models (PanEcho (Holste et al., 2025), EchoPrime (Vukadinovic et al., 2026)), and self-supervised representations (Image-MAE (Böhi et al., 2026), Video-LAMAE (Böhi et al., 2026), EchoJEPA-L (Munim et al., 2026)). For PanEcho and EchoPrime, we freeze their pretrained visual encoders and train new model-specific LVEF prediction heads on the training split of each benchmark. We include EchoJEPA-L as the published EchoJEPA baseline because its ViT-L visual encoder is comparable in scale to our DINOv3 ViT-L backbone. All direct evaluations report full test-set Mean Absolute

Error (MAE) and Mean Squared Error (MSE) with bootstrap standard errors across 10,000 video resamples using identical resampled indices. Downstream LVEF attention-probing protocol details and optimization schedules are listed in Appendix B.

For keyframe localization, we evaluate the frozen cyclic state readout (Section 3.4) using training-calibrated ED/ES phase anchors without updating model parameters or requiring ECG hardware gating. We report frame-level localization MAE, Root Mean Squared Error (RMSE), and R^2 against ground-truth clinical annotations using original-video frame indices. For the evaluated state-conditioned models, we report full-test estimates with bootstrap standard errors from 10,000 video resamples, using identical resampled indices across variants within each view.

4.2 Cyclic State Pretraining Enriches Visual Representations

We first evaluate whether pretraining the visual encoder E_ϕ under a state-conditioned predictive objective yields richer representations for downstream functional assessment. Table 1 compares frozen-backbone LVEF attention-probing performance across the EchoNet-Dynamic and EchoNet-Pediatric benchmarks.

Table 1: **Functional assessment from frozen visual encoders.** LVEF regression performance evaluated on EchoNet-Dynamic and EchoNet-Pediatric benchmarks. MAE is expressed in percentage points (%), and MSE in squared percentage points ($\%^2$). Direct evaluations report full-test estimates $\pm$ bootstrap standard errors from 10,000 video resamples. Published results (†) retain source-reported metrics; '–' denotes unreported values. Bold and underlined numbers mark the best and second-best results among directly evaluated models.

	EchoNet-Dynamic		EchoNet-Pediatric	
Method	MAE ↓	MSE ↓	MAE ↓	MSE ↓
General-purpose visual pretraining				
DINOv3 (Siméoni et al., 2025)	5.506 ± 0.141	55.776 ± 3.164	5.430 ± 0.191	61.724 ± 5.690
Task- and report-supervised echocardiography pretraining				
PanEcho (Holste et al., 2025)	<u>5.191</u>* ± 0.130	48.520 ± 2.915	4.877 ± 0.174	50.696 ± 5.752
EchoPrime (Vukadinovic et al., 2026)	5.292 ± 0.125	<u>48.471</u> ± 2.532	<u>4.807</u> ± 0.139	**40.309** ± 2.585
Self-supervised echocardiography pretraining				
Image-MAE† (Böhi et al., 2026)	8.14 ± 0.05	–	6.80 ± 0.04	–
Video-LAMAE† (Böhi et al., 2026)	6.78 ± 0.13	–	6.49 ± 0.09	–
EchoJEPA-L† (Munim et al., 2026)	5.76	–	5.12	–
CardiacState (ours)	**4.867** ± 0.121	**42.554** ± 2.249	**4.725** ± 0.175	<u>49.599</u> ± 5.188

* Note: PanEcho originally reported 5.53. To ensure a fair comparison, we froze the image encoder and fine-tuned only the task head, achieving better performance.

On EchoNet-Dynamic, CardiacState achieves the top performance across both metrics, reaching an MAE of $4.867 \pm 0.121\%$ and an MSE of $42.554 \pm 2.249\%^2$. Relative to the strongest task-supervised baseline PanEcho (MAE 5.191%), this represents a 0.324 percentage point (6.2%) error reduction, along with a 12.2% reduction in squared error relative to EchoPrime (MSE $48.471\%^2$). Compared to DINOv3 (5.506% MAE), CardiacState substantially improves downstream probe representations.

On EchoNet-Pediatric, performance between top methods is more nuanced. CardiacState achieves an MAE of $4.725 \pm 0.175\%$, which is on par with EchoPrime ($4.807 \pm 0.139\%$) and PanEcho ($4.877 \pm 0.174\%$), with the 0.082 point margin over EchoPrime falling well within standard error bounds (a statistical tie). On pediatric MSE, however, EchoPrime demonstrates lower dispersion ($40.309 \pm 2.585\%^2$) compared to CardiacState ($49.599 \pm 5.188\%^2$). This divergence between low MAE and higher MSE indicates that while CardiacState produces accurate median estimates, its predictions are subject to a small number of severe prediction outliers—likely driven by the pronounced anatomical variability and extreme heart rates characteristic of pediatric populations.

Because the downstream probing head operates on frozen features F_t without access to the state pathway during inference, these results confirm that forcing the predictor P_ω to forecast target features

through a bottlenecked cyclic state enriches E_ϕ with global physiological features that generalize effectively across adult and pediatric cohorts.

4.3 KEYFRAME LOCALIZATION FROM FROZEN CYCLIC STATES

We next evaluate whether the latent cardiac state vector $c_t \in S^1$ captures interpretable temporal dynamics without supervised fine-tuning or ECG signals. Using training-calibrated phase anchors ($\mathbf{a}_{\mathrm{ED}}, \mathbf{a}_{\mathrm{ES}}$) and cosine similarity matching (Equation 7), we localize keyframes across three standard acoustic views (A4C, PSAX, PLAX). Because each video provides one annotated ED and ES frame, we follow LMP (Yang et al., 2025) and compute MAE to the closest prediction.

Table 2: **Keyframe localization from frozen cyclic states.** LMP and OrbitalEcho provides the protocol-matched comparison, as all methods compute localization error to the temporally closest prediction when only one ED/ES pair is annotated. Other published baselines (†, transcribed from Chen et al. (2026)) are included as reference, as their prediction-matching procedures may differ or are not fully specified. CardiacState derives keyframes directly from its frozen state space $c_t \in S^1$ using cosine similarity to training-calibrated phase anchors. MAE and RMSE are measured in original-video frames; Avg denotes the unweighted mean across views. For CardiacState, $\pm$ denotes bootstrap standard error from 10,000 video resamples.

View	Method	MAE ↓	RMSE ↓	R^2 ↑
A4C	DeepPhase† (Farhad et al., 2023)	7.707	24.581	0.619
	OFM† (Lu et al., 2025)	2.832	<u>3.527</u>	<u>0.994</u>
	FrameONE (single)† (Chen et al., 2026)	2.883	3.980	0.993
	LMP (Yang et al., 2025)	2.5	–	–
	OrbitalEcho (Bonnici et al., 2026)	<u>2.245</u>	–	–
	CardiacState (ours)	**1.985** $\pm$ 0.024	**2.325** $\pm$ 0.022	**0.9967** $\pm$ 0.0002
PSAX	DeepPhase† (Farhad et al., 2023)	6.204	17.589	0.782
	OFM† (Lu et al., 2025)	2.970	<u>5.136</u>	0.978
	FrameONE (single)† (Chen et al., 2026)	<u>2.891</u>	5.274	<u>0.979</u>
	CardiacState (ours)	**2.027** $\pm$ 0.045	**2.452** $\pm$ 0.054	**0.9958** $\pm$ 0.0010
PLAX	DeepPhase† (Farhad et al., 2023)	7.204	20.828	0.678
	OFM† (Lu et al., 2025)	3.071	5.026	0.993
	FrameONE (single)† (Chen et al., 2026)	<u>2.509</u>	<u>4.150</u>	<u>0.996</u>
	CardiacState (ours)	**1.994** $\pm$ 0.049	**2.358** $\pm$ 0.046	**0.9962** $\pm$ 0.0006
Avg	DeepPhase† (Farhad et al., 2023)	7.038	20.999	0.693
	OFM† (Lu et al., 2025)	2.958	4.563	0.988
	FrameONE (single)† (Chen et al., 2026)	2.761	<u>4.468</u>	<u>0.989</u>
	LMP (Yang et al., 2025)	2.5	–	–
	OrbitalEcho (Bonnici et al., 2026)	<u>2.245</u>	–	–
	CardiacState (ours)	**2.002** $\pm$ 0.023	**2.378** $\pm$ 0.024	**0.9962** $\pm$ 0.0004

As shown in Table 2, CardiacState achieves an overall average MAE of 2.002 ± 0.023 frames across all views (1.985 ± 0.024 in A4C, 2.027 ± 0.045 in PSAX, and 1.994 ± 0.049 in PLAX), with an inter-view range of 0.042 frames. Corresponding RMSE values are 2.325 ± 0.022, 2.452 ± 0.054, and 2.358 ± 0.046 frames, respectively; the mean across views is 2.378 ± 0.024 frames.

Under the same closest-prediction evaluation used by LMP and OrbitalEcho, CardiacState reduces A4C mean localization MAE from 2.245 to 1.985 ± 0.024 frames. Results from other supervised detectors are included as context reference. Crucially, CardiacState performs keyframe localization by querying its frozen, self-supervised cardiac state space S^1 with training-calibrated phase anchors, without training a task-specific detection head.

These low localization errors occur across datasets with distinct acquisition frame-rate distributions (Appendix F), suggesting robustness to the frame-rate differences observed here. We think that sampling pretraining clips at multiple frame gaps and explicitly conditioning on elapsed physical time contribute to this robustness.

4.4 Ablation Studies

To isolate the effects of cardiac progression regularization and the temporal bottleneck mask, we evaluate three pretraining variants that all retain the cardiac phase prediction module. At $K = 4$, we compare models without and with $\mathcal{L}_{\text{cardiac}}$, both giving the predictor the inferred state c_t and four-frame visual context. The full CardiacState model retains the loss but restricts direct visual context to the current frame F_t ($K = 1$), so historical context reaches the predictor through c_t.

Table 3: **Ablation of cardiac progression regularization and temporal bottleneck.** All three variants retain the cardiac phase prediction module. Frozen-encoder LVEF estimation and keyframe localization from frozen cyclic states are evaluated across the matched pretraining variants. MAE and MSE are in percentage points (%) and squared percentage points ($\%^2$). Localization estimates are reported with bootstrap standard errors from 10,000 video resamples. Bold and underlined values mark the top two results; Appendix D reports per-view localization.

	EchoNet-Dynamic		EchoNet-Pediatric		Localization
Method	MAE ↓	MSE ↓	MAE ↓	MSE ↓	Mean MAE (frames) ↓
$K = 4$, w/o $\mathcal{L}_{\text{cardiac}}$	5.204 ± 0.133	50.154 ± 2.980	5.070 ± 0.191	58.648 ± 6.110	18.396 ± 0.570
$K = 4$, w/ $\mathcal{L}_{\text{cardiac}}$	$\underline{5.028} \pm 0.124$	$\underline{45.549} \pm 2.385$	$\underline{4.811} \pm 0.175$	$\underline{50.835} \pm 5.397$	$\underline{2.107} \pm 0.027$
$K = 1$, w/ $\mathcal{L}_{\text{cardiac}}$	**4.867** ± 0.121	**42.554** ± 2.249	**4.725** ± 0.175	**49.599** ± 5.188	**2.002** ± 0.023

Cardiac Progression Regularization Organizes the Learned State. At fixed $K = 4$, all components of the cardiac phase prediction module are present in both variants. Adding $\mathcal{L}_{\text{cardiac}}$ reduces mean keyframe localization MAE from 18.396 to 2.107 frames (Table 3). It also reduces LVEF estimation MAE from 5.204% to 5.028% on EchoNet-Dynamic and from 5.070% to 4.811% on EchoNet-Pediatric. Thus, cyclic parameterization and phase-conditioned prediction alone do not yield a comparably useful cardiac coordinate in this setting; the progression loss is important for organizing the learned state and improving the frozen visual representations.

Strict Bottlenecking Strengthens Cardiac State Conditioning. Holding $\mathcal{L}_{\text{cardiac}}$ present, enforcing the strict temporal bottleneck (M_1, $K = 1$) further reduces keyframe localization error from 2.107 ± 0.027 to 2.002 ± 0.023 frames and improves LVEF estimation MAE from 5.028% to 4.867% on EchoNet-Dynamic and from 4.811% to 4.725% on EchoNet-Pediatric. With $K = 1$, the predictor can access visual history only through the S^1 cardiac state vector c_t. The additional gains are consistent with stronger coupling between physiological state inference and future-feature prediction when direct multi-frame visual context is removed.

A qualitative visualization of the corresponding learned state trajectories is provided in Appendix E. For a representative video, the visualization shows that removing $\mathcal{L}_{\text{cardiac}}$ produces a near-stationary state, whereas introducing the progression regularizer and enforcing the $K = 1$ temporal bottleneck yield increasingly well-organized forward trajectories on S^1.

5 Discussion & Limitations

Improved frozen-backbone LVEF estimation on EchoNet-Dynamic and EchoNet-Pediatric, together with ED/ES localization across three acoustic views, supports an inferred "action-like" cardiac state as a useful conditioning signal for echocardiographic video. The LVEF probe never sees the predicted cardiac state c_t, so its gains suggest that state-conditioned prediction enriches the visual backbone with information about ventricular change; keyframe localization, in turn, reads the frozen state directly. Ablations show that cardiac progression regularization drives the large localization improvement, while the strict $K = 1$ bottleneck adds a further gain. These results suggest that the S^1 parameterization works best when organized by the progression objective and coupled to a strict temporal bottleneck. **Future work.** Although the formulation permits variable angular velocity and heart-rate-dependent timing, arrhythmias or ectopic beats may violate a coherent recurrent trajectory, reducing the reliability of a single cyclic coordinate and fixed phase anchors. Future work should evaluate such rhythms and explore event-aware or piecewise-cyclic states. Anchor dependence could also be reduced through anatomy-based alignment.

Ethics Statement

This study develops representations for retrospective echocardiographic analysis using a private pretraining dataset and public downstream benchmarks. The pretraining dataset was collected with institutional review board approval and a waiver of informed consent for retrospective analysis of existing deidentified clinical data. This private pretraining dataset does not overlap with the downstream benchmarks. Public datasets are used under their respective data-use agreements.

Reproducibility Statement

Section 3 defines the architecture and learning objective. Section 4 specifies the datasets and comparisons, and Appendices A–E give architectural dimensions, optimization, calibration, evaluation, per-view ablations, and qualitative cardiac-state trajectory analysis. We plan to release code, configurations, and evaluation summaries subject to data-use permissions.

AI Use Statement

Generative AI tools assisted with manuscript restructuring, language editing, LaTeX drafting, and figure-code preparation under author supervision. The authors are responsible for the scientific claims, experimental design, numerical results, citation verification, and final manuscript.

References

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.

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.

Simon Böhi, Irene Cannistraci, Sergio Muñoz Gonzalez, Moritz Vandenhirtz, Sonia Laguna, Samuel Ruiperez-Campillo, Max Krähenmann, Andrea Agostini, Ece Ozkan, Thomas M Sutter, et al. Beyond independent frames: Latent attention masked autoencoders for multi-view echocardiography. *arXiv preprint arXiv:2604.15096*, 2026.

John Bonnici, Matthew Baugh, Aleksandra Kulbaka, Sarah Cechnicka, Bernhard Kainz, and Alberto Gomez. Self-supervised cardiac phase detection via single-parameter latent orbits. *arXiv preprint arXiv:2609.11650*, 2026.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Rusi Chen, Yuhao Huang, Hongyuan Zhang, Chao Tian, Shunan Ji, Yuhan Zhang, and Dong Ni. Frameone: Hierarchical motion modeling for universal multi-view echocardiographic keyframe detection. *arXiv preprint arXiv:2607.00748*, 2026.

Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5):1481–1488, 2024.

Charles S Chung, Mustafa Karamanoglu, and Sándor J Kovács. Duration of diastole and its phases as a function of heart rate during supine bicycle exercise. *American Journal of Physiology-Heart and Circulatory Physiology*, 287(5):H2003–H2008, 2004.

Grant Duffy, Paul P Cheng, Neal Yuan, Bryan He, Alan C Kwan, Matthew J Shun-Shin, Kevin M Alexander, Joseph Ebinger, Matthew P Lungren, Florian Rader, et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA cardiology*, 7(4):386–395, 2022.

Ashley Edwards, Himanshu Sahni, Yannick Schroecker, and Charles Isbell. Imitating latent policies from observation. In *International conference on machine learning*, pp. 1755–1763. PMLR, 2019.

Moomal Farhad, Mohammad Mehedy Masud, Azam Beg, Amir Ahmad, Luai A Ahmed, and Sehar Memon. Cardiac phase detection in echocardiography using convolutional neural networks. *Scientific reports*, 13(1):8908, 2023.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Gregory Holste, Evangelos K Oikonomou, Márton Tokodi, Attila Kovács, Zhangyang Wang, and Rohan Khera. Complete ai-enabled echocardiography interpretation with multitask deep learning. *Jama*, 334(4):306–318, 2025.

Elisabeth S Lane, Neda Azarmehr, Jevgeni Jevsikov, James P Howard, Matthew J Shun-Shin, Graham D Cole, Darrel P Francis, and Massoud Zolgharni. Multibeat echocardiographic phase detection using deep neural networks. *Computers in Biology and Medicine*, 133:104373, 2021.

Roberto M Lang, Luigi P Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A Flachskampf, Elyse Foster, Steven A Goldstein, Tatiana Kuznetsova, et al. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging. *European Heart Journal-Cardiovascular Imaging*, 16(3):233–271, 2015.

Fabian Laumer, Gabriel Fringeli, Alina Dubatovka, Laura Manduchi, and Joachim M Buhmann. Deepheartbeat: Latent trajectory learning of cardiac cycles using cardiac ultrasounds. In *Machine Learning for Health*, pp. 194–212. PMLR, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yuhuan Lu, Guanghua Tan, Bin Pu, Pak-Hei Yeung, Hang Wang, Shengli Li, Jagath C Rajapakse, and Kenli Li. Optical flow-enhanced mamba u-net for cardiac phase detection in ultrasound videos. *IEEE Transactions on Medical Imaging*, 44(7):2831–2841, 2025.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Alif Munim, Adibvafa Fallahpour, Teodora Szasz, Ahmadreza Attarpour, River Jiang, Brana Sooriyakanthan, Maala Sooriyakanthan, Heather Whitney, Jeremy Slivnick, Barry Rubin, et al. Echojepa: A latent predictive foundation model for echocardiography. *arXiv preprint arXiv:2602.02603*, 2026.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 6960–6970. IEEE, 2021.

Ashwath Radhachandran, Vedrana Ivezić, Shreeram Athreya, Ronit Anilkumar, Corey W Arnold, and William Speier. Us-jepa: A joint embedding predictive architecture for medical ultrasound. *arXiv preprint arXiv:2602.19322*, 2026.

Charitha D Reddy, Leo Lopez, David Ouyang, James Y Zou, and Bryan He. Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. *Journal of the American Society of Echocardiography*, 36(5):482–489, 2023.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *International Conference on Learning Representations*, volume 2024, pp. 9379–9395, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Milos Vukadinovic, I-Min Chiu, Xiu Tang, Neal Yuan, Tien-Yu Chen, Paul Cheng, Debiao Li, Susan Cheng, Bryan He, and David Ouyang. Comprehensive echocardiogram evaluation with view primed vision language ai. *Nature*, 650(8103):970–977, 2026.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Guo Guo. Aurora: Towards universal generative multimodal time series forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 103239–103277, 2026.

Yingyu Yang, Qianye Yang, Kangning Cui, Can Peng, Elena D'Alberti, Netzahualcoyotl Hernandez-Cruz, Olga Patey, Aris T Papageorghiou, and J Alison Noble. Latent motion profiling for annotation-free cardiac phase detection in adult and fetal echocardiography videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 316–325. Springer, 2025.

Ziyang Zhang, Qinxin Wu, Sirui Ding, Xiaolong Wang, and Jiancheng Ye. Echo-vision-fm: a pre-training and fine-tuning framework for echocardiogram video vision foundation model. *Nature Communications*, 17(1):19, 2025.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. Astra: General interactive world model with autoregressive denoising. In *International Conference on Learning Representations*, volume 2026, pp. 79167–79184, 2026.

A Implementation Details

Table 4 summarizes the architecture used for pretraining and the frozen-backbone attention probe.

Table 4: **Architecture configuration.** Dimensions and settings for the visual encoder, cyclic-state pathway, future predictor, and LVEF attention probe.

Component	Configuration
Visual encoder	DINOv3 ViT-L; 256×256 input frames
Feature projections	Two-layer MLPs; hidden width 2048, output width 1024; GELU and Batch Normalization
State encoder	Four learned queries; two causal cross-attention blocks; four heads; width 1024; dropout 0.1
State decoder	$1024 \to 512 \to 2$ MLP; SiLU and unit normalization
State conditioning	$2 \to 128 \to 1024$ MLP
Time conditioning	512 sinusoidal frequency pairs; maximum period 4,000 ms; $1024 \to 4096 \to 1024$ MLP
Future predictor	40 blocks; 24 heads of dimension 64; residual width 1024; feed-forward expansion four; dropout 0.1
EF attention probe	16 uniformly sampled frames; four self-attention blocks with eight heads; final Layer Normalization and linear regression

Visual encoder and projections. The visual encoder is initialized from DINOv3. Its normalized class token passes through the encoder projection to produce $d = 1024$ predictive features. The predictor output uses a separate projection with the same MLP dimensions. During training, synchronized Batch Normalization aggregates over batch and frame axes; at inference, it uses running statistics.

State module and conditioning. Each learned query receives a temporal positional embedding. Query outputs are averaged before decoding and unit normalization. Learned scalar weights combine the state and elapsed-time branches before Layer Normalization.

Predictor. The attention layer projects into a 1,536-dimensional concatenated head space and then back to the residual width. The causal mask restricts each temporal position to its current visual token. The fused condition supplies Adaptive Layer Normalization scale, shift, and gate parameters initialized at zero.

EF token aggregation. For sampled original frame indices $i_1, \ldots, i_{16}$, the temporal coordinates are $\widetilde{i}_t = 15(i_t - i_{\min})/(i_{\max} - i_{\min} + 10^{-6})$. The readout combines these coordinates with patch-grid coordinates in multidimensional rotary position encoding. Patch tokens from all sampled frames are concatenated with a global token initialized to the mean of frame class tokens. The attention blocks and final Layer Normalization update the sequence; a linear regressor maps the global token to EF. The architecture is summarized in Table 4. Training uses squared error, and checkpoint selection within each run uses validation MAE.

B Optimization and Statistical Reporting

Pretraining. Our pretraining configuration uses eight NVIDIA H200 GPUs, bfloat16 precision, and 16 videos per GPU, giving a global batch of 128. Training lasts ten epochs. AdamW uses weight decay 10^{-3} and gradient-norm clipping at 1.0 (Loshchilov & Hutter, 2017). Newly initialized modules use a peak learning rate of 10^{-4}, a one-epoch warmup, and cosine decay to 10^{-6}. The encoder learning rate is zero for two epochs, then follows a delayed cosine schedule from 5×10^{-6} to 5×10^{-7}. Gradients from the state branch into its input features are stopped during the same initial interval. Frames are resized with aspect ratio preserved, padded to 256×256, rotated by up to $15°$, and ImageNet normalized. Appendix A records the remaining dimensions and regularization settings.

EF readout fitting. The downstream configuration uses two GPUs, 32 videos per device, 30 epochs, and three warmup epochs. AdamW minimizes squared EF error with weight decay 10^{-3} and a cosine learning-rate schedule. The CardiacState configuration uses a four-block readout with relative frame-position encoding and learning rates of 2×10^{-4} for Dynamic and 5×10^{-4} for Pediatric. The hyperparameter search covers learning rates $\{5 \times 10^{-5}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$, depths $\{4, 6\}$, and absolute versus normalized relative temporal encodings. Within each run, the checkpoint with the lowest validation MAE is saved for test evaluation.

Metrics and uncertainty. EF MAE and MSE are computed on the full test predictions in percentage points and squared percentage points. For directly evaluated models, $\pm$ denotes bootstrap standard error from 10,000 resamples of video rows, using the same resampled indices across models within each dataset. Published references retain their source uncertainty. Localization uses original-video frames and reports an arithmetic mean over the three views. Localization uncertainty uses the same video-level bootstrap strategy, with identical resampled video indices across model variants within each view; the reported Avg uncertainty is the bootstrap standard error of the view-averaged metric. The duration study evaluates checkpoints after three, six, and ten epochs with the same sliding-window state aggregation and anchor-based readout.

C Full-Video Keyframe Evaluation

Sliding-window inference and state aggregation. We evaluate 16-frame observation windows sampled at a stride (gap) of 2 original frames, with an 8-frame overlap between consecutive windows. For a video frame i covered by overlapping windows, each window $w \in \mathcal{W}_i$ predicts a unit state vector $\mathbf{c}_i^{(w)} \in S^1 \subset \mathbb{R}^2$ based on its specific context. We aggregate these multi-window predictions by linear vector averaging:

$$\bar{\mathbf{c}}_i = \frac{1}{|\mathcal{W}_i|} \sum_{w \in \mathcal{W}_i} \mathbf{c}_i^{(w)}. \tag{8}$$

This operation averages two-dimensional vector components directly rather than wrapped scalar phase angles. Frames covered by a single window retain that individual vector estimate. Calibration and query videos undergo identical aggregation; for anchor calibration, we additionally normalize each nonzero average as $\widetilde{\mathbf{c}}_i = \bar{\mathbf{c}}_i / \|\bar{\mathbf{c}}_i\|_2$.

Anchor calibration. For each dataset and model checkpoint, let $\mathcal{A}_k$ denote the set of ground-truth annotated frames for phase $k \in \{\mathrm{ED}, \mathrm{ES}\}$ in the calibration split. We compute the aggregated state vector $\bar{\mathbf{c}}_i$ for each annotated frame using Equation 8 and define the dataset-level phase anchor $\mathbf{a}_k$ via circular averaging of the normalized vectors:

$$\mathbf{a}_k = \frac{\sum_{i \in \mathcal{A}_k} \widetilde{\mathbf{c}}_i}{\left\| \sum_{i \in \mathcal{A}_k} \widetilde{\mathbf{c}}_i \right\|_2}. \tag{9}$$

All pretrained visual encoder and predictor parameters remain strictly frozen during this calibration step.

Inference and event extraction. During inference on a query clip, frame-level phase scores are calculated as $S_{\mathrm{ED}}(i) = \bar{\mathbf{c}}_i^\top \mathbf{a}_{\mathrm{ED}}$ and $S_{\mathrm{ES}}(i) = \bar{\mathbf{c}}_i^\top \mathbf{a}_{\mathrm{ES}}$. Each is the average of the per-window cosine similarities because the window state vectors and anchors are unit-normalized; it is not the cosine similarity of the unnormalized average $\bar{\mathbf{c}}_i$. To identify phase-dominant regions, we compute the End-Diastolic probability score $p(i)$:

$$p(i) = \sigma\left(\frac{S_{\mathrm{ED}}(i) - S_{\mathrm{ES}}(i) + b}{T_{\mathrm{score}}} \right), \tag{10}$$

where $\sigma(\cdot)$ is the standard logistic sigmoid function, $T_{\mathrm{score}} = 1.0$ is the score temperature, and $b = 0.0$ is the score bias. Thresholding at $p(i) \geq 0.5$ (equivalent to $S_{\mathrm{ED}}(i) \geq S_{\mathrm{ES}}(i)$) segments the sequence into ED-dominant regions, while $p(i) < 0.5$ defines ES-dominant regions. Within each ED segment, the frame maximizing $S_{\mathrm{ED}}(i)$ is selected as the ED keyframe; within each ES segment, the frame maximizing $S_{\mathrm{ES}}(i)$ is selected as the ES keyframe, evaluated using a 5-frame local temporal window with ties resolved toward the segment center.

D PER-VIEW LOCALIZATION ABLATIONS

Table 5 reports localization metrics for the three phase-module configurations in Section 4.4. At fixed $K = 4$, adding $\mathcal{L}_{\text{cardiac}}$ reduces MAE in every acoustic view; with the loss present, restricting predictor context from $K = 4$ to $K = 1$ reduces MAE further in every view. All variants use the same encoder initialization, pretraining data, optimization schedule, and downstream evaluation protocol.

Table 5: **Per-view effects of cardiac progression regularization and predictor visual context.** All three variants retain the cardiac phase prediction module. MAE and RMSE are in original-video frames. The first two variants isolate $\mathcal{L}_{\text{cardiac}}$ at $K = 4$; the last two compare $K = 4$ and $K = 1$ with the loss present. Estimates are reported with bootstrap standard errors from 10,000 video resamples. Avg averages the three displayed views. Bold values mark the best result for each metric and view.

View	Method	MAE ↓	RMSE ↓	R^2 ↑
A4C	$K = 4$, w/o $\mathcal{L}_{\text{cardiac}}$	28.475 ± 0.997	49.082 ± 1.395	-0.429 ± 0.094
	$K = 4$, w/ $\mathcal{L}_{\text{cardiac}}$	2.105 ± 0.029	2.518 ± 0.028	0.9962 ± 0.0002
	$K = 1$, w/ $\mathcal{L}_{\text{cardiac}}$	**1.985** ± 0.024	**2.325** ± 0.022	**0.9967** ± 0.0002
PSAX	$K = 4$, w/o $\mathcal{L}_{\text{cardiac}}$	15.183 ± 0.883	27.059 ± 1.690	0.501 ± 0.137
	$K = 4$, w/ $\mathcal{L}_{\text{cardiac}}$	2.083 ± 0.047	2.573 ± 0.051	0.9953 ± 0.0011
	$K = 1$, w/ $\mathcal{L}_{\text{cardiac}}$	**2.027** ± 0.045	**2.452** ± 0.054	**0.9958** ± 0.0010
PLAX	$K = 4$, w/o $\mathcal{L}_{\text{cardiac}}$	11.531 ± 1.055	24.956 ± 2.077	0.572 ± 0.101
	$K = 4$, w/ $\mathcal{L}_{\text{cardiac}}$	2.135 ± 0.061	2.597 ± 0.058	0.9954 ± 0.0008
	$K = 1$, w/ $\mathcal{L}_{\text{cardiac}}$	**1.994** ± 0.049	**2.358** ± 0.046	**0.9962** ± 0.0006
Avg	$K = 4$, w/o $\mathcal{L}_{\text{cardiac}}$	18.396 ± 0.570	33.699 ± 1.011	0.215 ± 0.065
	$K = 4$, w/ $\mathcal{L}_{\text{cardiac}}$	2.107 ± 0.027	2.563 ± 0.028	0.9956 ± 0.0005
	$K = 1$, w/ $\mathcal{L}_{\text{cardiac}}$	**2.002** ± 0.023	**2.378** ± 0.024	**0.9962** ± 0.0004

The per-view breakdown in Table 5 first isolates the progression loss at fixed $K = 4$: adding $\mathcal{L}_{\text{cardiac}}$ reduces MAE from 28.475 to 2.105 frames in Apical Four-Chamber (A4C), from 15.183 to 2.083 in Parasternal Short-Axis (PSAX), and from 11.531 to 2.135 in Parasternal Long-Axis (PLAX). With the loss present, restricting predictor context from $K = 4$ to $K = 1$ yields further reductions in every view (A4C: 2.105 to 1.985; PSAX: 2.083 to 2.027; PLAX: 2.135 to 1.994), lowering the overall average to 2.002 frames (2.378 RMSE). The consistent, smaller second improvement supports the benefit of routing visual history through the S^1 state, although these results do not directly measure the predictor's use of visual persistence shortcuts.

E QUALITATIVE ANALYSIS OF LEARNED CARDIAC-STATE DYNAMICS

To qualitatively examine how the cardiac progression regularizer and temporal bottleneck shape the learned cyclic representation, we visualize the inferred cardiac-state trajectory for the same representative video under three pretraining variants in Figure 2. All three retain the cardiac phase prediction module; the top variant omits only $\mathcal{L}_{\text{cardiac}}$. For each frame, we convert the predicted unit state $c_t \in S^1$ to its corresponding angular coordinate θ_t and plot the wrapped angle over time. Transitions from 2π back to 0 therefore indicate the natural wrap-around of the circular state space rather than backward motion.

Without $\mathcal{L}_{\text{cardiac}}$, the inferred state remains confined to a narrow angular range and exhibits little systematic temporal progression. Although the model retains a two-dimensional cyclic-state parameterization, the representation does not spontaneously organize itself into a consistently advancing cardiac coordinate. This qualitative behavior is consistent with the substantial degradation in keyframe localization observed for the corresponding ablation in Table 5.

Introducing $\mathcal{L}_{\text{cardiac}}$ produces a markedly different trajectory: the state repeatedly traverses the circular manifold and wraps from 2π to 0, indicating that temporal progression has become organized along the learned cyclic coordinate.

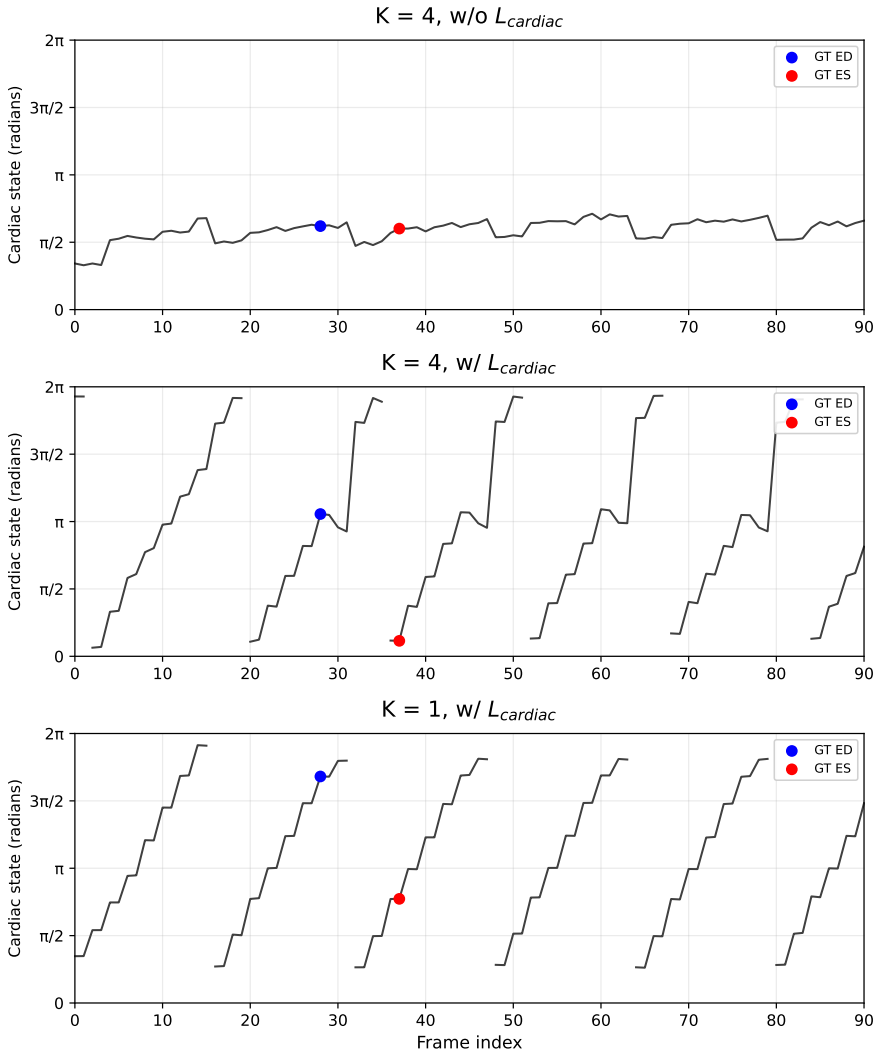


Figure 2: **Qualitative comparison of learned cardiac-state trajectories.** The inferred cardiac-state angle is plotted over frame index for the same representative video under three pretraining configurations, all of which retain the cardiac phase prediction module. **Top:** removing $\mathcal{L}_{\text{cardiac}}$ yields a largely stationary trajectory without consistent forward cyclic progression. **Middle:** adding $\mathcal{L}_{\text{cardiac}}$ with $K = 4$ organizes the state into repeated traversals of S^1, although local irregularities remain. **Bottom:** the full CardiacState model with the strict $K = 1$ temporal bottleneck exhibits a more consistently forward-moving trajectory between circular wrap-around events. Blue and red markers denote the ground-truth ED and ES frames, respectively. The absolute angular origin is not aligned across models; the visualization is intended to compare temporal organization and cyclic progression rather than absolute phase angle.

With $K = 4$, however, local stalls and irregular progression remain visible within individual cycles. Under the full CardiacState configuration ($K = 1$), the trajectory becomes more consistently forward-moving between successive wrap-around events. This behavior is consistent with the temporal-bottleneck hypothesis in Section 4.4: when direct multi-frame visual context is removed, historical dynamics must be communicated through the cardiac state, encouraging a cleaner representation of continuous physiological progression.

This representative trajectory is consistent with the quantitative ablations: $\mathcal{L}_{\text{cardiac}}$ is important for establishing forward cyclic organization, while the strict $K = 1$ bottleneck further regularizes how temporal information is represented on S^1.

F Acquisition Frame-Rate Distributions

Figure 3 summarizes the acquisition frame rates of the three downstream evaluation datasets, with percentages normalized separately for each dataset. EchoNet-LVH is concentrated near 20 FPS, EchoNet-Dynamic near 50 FPS, and EchoNet-Pediatric has a broader distribution centered near 50 FPS with a high-frame-rate tail extending beyond 100 FPS. Thus, the benchmarks cover distinct acquisition timing regimes. For a fixed gap of two original-video frames in keyframe evaluation, the elapsed physical time between sampled observations varies with the video frame rate. Together with the low ED/ES localization errors across all three datasets in Tables 2 and 5, these distributions suggest that CardiacState remains effective across the observed frame-rate differences. We hypothesize that sampling pretraining clips at multiple frame gaps and explicitly conditioning on elapsed physical time contribute to this robustness.

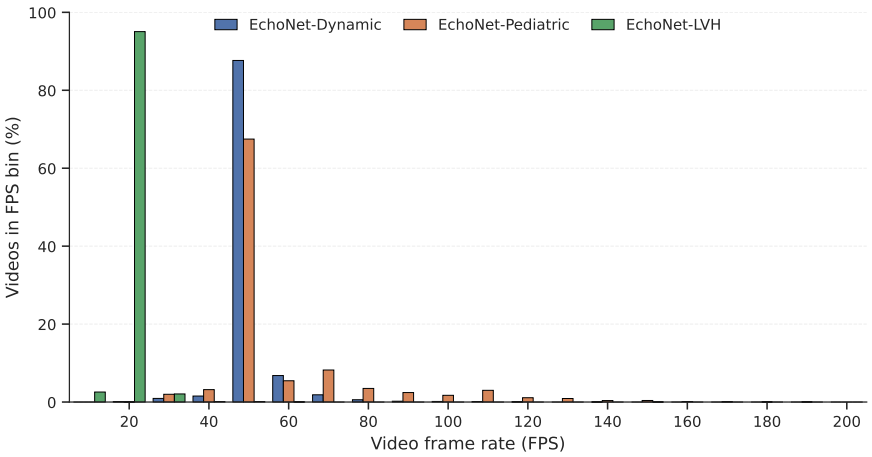


Figure 3: **Acquisition frame-rate distributions of the downstream evaluation datasets.** Each bar shows the percentage of videos in an FPS bin within its dataset; percentages are normalized separately for EchoNet-Dynamic, EchoNet-Pediatric, and EchoNet-LVH. EchoNet-LVH is concentrated at lower frame rates, whereas EchoNet-Pediatric spans the broadest range.