Self-supervised Learning of Echocardiographic Video Representations via Online Cluster Distillation

Divyanshu Mishra[1] **Mohammadreza Salehi**[4] **Pramit Saha**[1] **Olga Patey**[2]
Aris T. Papageorghiou[2] **Yuki M. Asano**[3] **J. Alison Noble**[1]
[1]Department of Engineering Science, University of Oxford
[2]Nuffield Department of Women's and Reproductive Health, University of Oxford
[3]Fundamental AI Lab, University of Technology Nuremberg
[4] University of Amsterdam
divyanshu.mishra@eng.ox.ac.uk

Abstract

Self-supervised learning (SSL) has achieved major advances in natural images and video understanding, but challenges remain in domains like echocardiography (heart ultrasound) due to subtle anatomical structures, complex temporal dynamics, and the current lack of domain-specific pre-trained models. Existing SSL approaches such as contrastive, masked modeling, and clustering-based methods struggle with high intersample similarity, sensitivity to low PSNR inputs common in ultrasound, or aggressive augmentations that distort clinically relevant features. We present DISCOVR (Distilled Image Supervision for Cross Modal Video Representation), a self-supervised dual branch framework for cardiac ultrasound video representation learning. DISCOVR combines a clustering-based video encoder that models temporal dynamics with an online image encoder that extracts fine-grained spatial semantics. These branches are connected through a semantic cluster distillation loss that transfers anatomical knowledge from the evolving image encoder to the video encoder, enabling temporally coherent representations enriched with fine-grained semantic understanding.Evaluated on six echocardiography datasets spanning fetal, pediatric, and adult populations, DISCOVR outperforms both specialized video anomaly detection methods and state-of-the-art video-SSL baselines in zero-shot and linear probing setups,achieving superior segmentation transfer and strong downstream performance on clinically relevant tasks such as LVEF prediction. Code available at: https://github.com/mdivyanshu97/DISCOVR

1 Introduction

Modeling dynamic content in video data presents significant challenges due to complex spatio-temporal relationships, high redundancy between frames, and the need to capture both short- and long-range temporal dependencies [41, 32]. Echocardiography (heart or cardiac ultrasound) exemplifies these video understanding challenges [23, 22]. With high frame rates (30–80 fps) [21], complex anatomical motion, and variability in image appearance caused by speckle, shadowing artifacts, and ultrasound probe variability [18], automated echocardiography analysis requires sophisticated temporal modeling approaches. The information density in these videos is high, where features critical for diagnosis may appear as subtle variations in wall motion, valve function, or blood flow patterns that manifest only when viewed dynamically across multiple frames. Moreover, the appearance of the heart can change drastically across different cardiac views, patient populations, and imaging equipment. Developing robust video SSL models for comprehensive video understanding in echocardiography faces additional obstacles due to data limitations. Expert annotations are costly, labor-intensive, and if based on real-world hospital data often incomplete, capturing only specific

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-2045

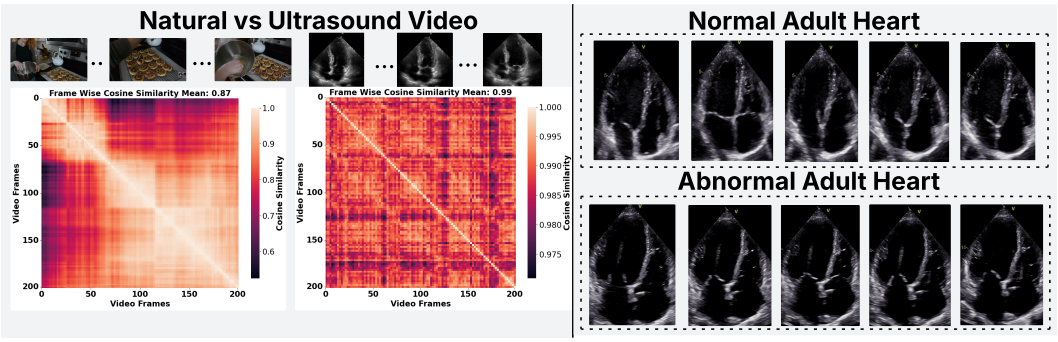

Figure 1: Figure (left) compares two fine-grained videos: a natural scene of a person baking (left) and an adult fetal heart ultrasound (right). The frame-level cosine similarity matrix, computed using a pretrained VideoMAE model, shows that ultrasound frames are highly similar (mean=0.99), with only minor local variations. This highlights the difficulty in distinguishing individual frames in such medical videos. Figure (right) compares normal and abnormal adult echocardiograms that appear nearly identical. However, on close inspection, it is revealed that the abnormal heart shows severe biventricular systolic dysfunction and a dilated, globular left ventricle, underscoring the subtlety of cardiac defects and the need for fine-grained structural analysis.

aspects of the rich information contained in these videos. This scarcity of labeled data motivates SSL approaches that can leverage abundant unlabeled echocardiograms for model development [32, 24, 6].

Several SSL frameworks have been proposed for learning meaningful video representations, each with particular limitations in the echocardiography context. Masked video modeling methods [35, 10, 9] tend to focus on reconstructing low-level image features like textures or edges, limiting their ability to capture high-level semantic information critical for clinical interpretation. This is especially problematic for ultrasound, which inherently exhibits a low signal-to-noise ratio (SNR), making approaches that rely on low-level pixel representations ineffective. Contrastive learning methods [13, 27] struggle due to high inter-sample similarity and limited effective augmentations, making it difficult to construct informative positive and negative pairs, often leading to representation collapse. Clustering-based SSL methods have demonstrated strong semantic learning through self-distillation but rely heavily on aggressive augmentations that risk disrupting essential anatomical details required for fine-grained understanding.

To address these limitations, we propose DISCOVR (*Distilled Image Supervision for Cross-Modal Video Representation*), a dual branch SSL framework tailored for echocardiography that jointly captures temporal dynamics and fine-grained semantic structure. The video encoder is trained to model temporal features using a clustering-based objective applied to masked video tokens, while an online image encoder separately learns spatially rich and anatomically meaningful representations from masked image views. To bridge the gap between spatial and temporal learning, we introduce a semantic cluster distillation loss that transfers knowledge from the evolving image encoder to the video encoder through semantic cluster alignment. This enables the video encoder to embed fine-grained semantic detail into its temporally coherent representations, without relying on pretrained models or heavy augmentations.

We extensively evaluate DISCOVR on six echocardiography datasets that span fetal, pediatric, and adult populations, covering anomaly detection, classification (linear probing and zero-shot transfer), and segmentation tasks. DISCOVR consistently outperforms prior self-supervised and anomaly detection methods. It achieves an average F1 improvement of 3.4% for anomaly detection, a 2.4% gain in linear probing, and a 1.5% increase in balanced accuracy under zero shot evaluation. For segmentation, DISCOVR delivers a 3.1% relative improvement in Dice score (from 81.9 to 84.4), despite using a simple segmentation head compared to more complex baseline architectures. These results demonstrate that integrating spatial semantics with temporal dynamics through cross-modal distillation yields robust and generalizable cardiac ultrasound video representations.

Overall, our contibutions are as follows:

- We develop an SSL method that jointly models temporal dynamics and spatial semantics by integrating video self-distillation with an evolving semantic image encoder, without labels, pretrained models, or augmentations.

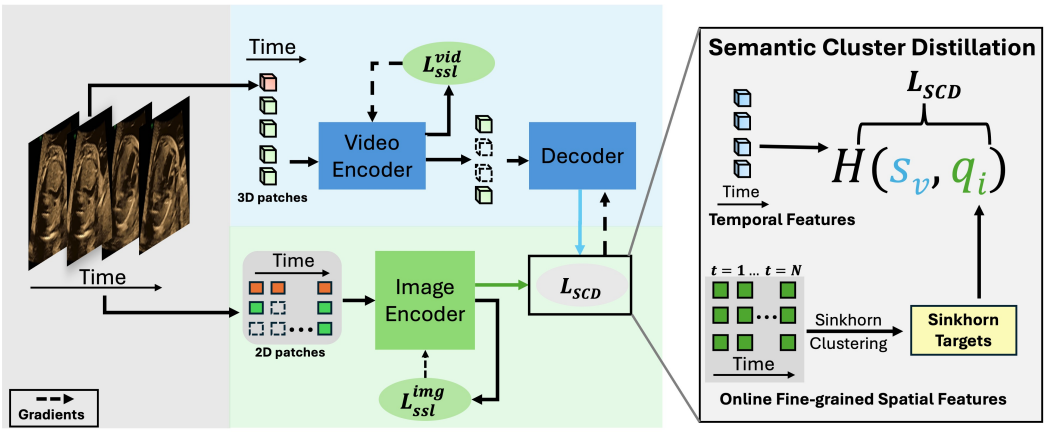

Figure 2: **Overview of the DISCOVR framework.** An input video is tokenized into 3D patches for the video branch and per-frame 2D patches for the image branch. Both encoders perform masked self-distillation. Masked video tokens are reconstructed by the video decoder, and dense semantic features are extracted from the image encoder. The $\mathcal{L}_{\text{SCD}}$ loss then aligns these outputs, distilling fine-grained spatial semantics into the video representation to produce rich spatio-temporal features.

- We introduce a novel online semantic distillation loss that continually transfers anatomical knowledge from the evolving image encoder to the video encoder, enriching its temporal representations with fine-grained spatial semantics to better capture clinically relevant spatio-temporal patterns in echocardiography.
- DISCOVR is, to our knowledge, the most comprehensive self-supervised video representation model for echocardiography to date. Trained solely on normal videos, it models healthy heart dynamics and detects pathology as deviations, eliminating the need for labeled abnormal cases. Evaluated across six datasets spanning fetal, pediatric, and adult cohorts, DISCOVR demonstrates strong generalization in zero-shot classification, linear probing, anomaly detection, and segmentation, and achieves state-of-the-art performance on downstream cardiac function estimation (LVEF prediction), making it a versatile backbone for ultrasound analysis.

2 Related Work

Self-supervised learning (SSL) aims to learn feature extractors directly from raw data by solving an intrinsic task using supervision signals derived from the data itself, eliminating the need for manual labels. Early image-based SSL relied on handcrafted pretext tasks such as solving jigsaw puzzles [26], predicting rotations [11], or colorizing grayscale inputs [45]. Recent methods have shifted towards instance discrimination via contrastive learning [42, 13, 5]. To understand how these ideas extend to video and medical domains, we review the most relevant self-supervised methods in both areas, highlighting shared limitations and how DISCOVR addresses them.

Video Self-Supervised Learning. Extending SSL to video introduces additional temporal complexity, inspiring tasks such as frame order prediction [25, 43], spatio-temporal jigsaws [16], and playback pace prediction [3, 38]. Recently, masked video modeling has become the dominant approach: VideoMAE [35] reconstructs raw pixels from masked tubelets using a ViT backbone. MGMAE [14] predicts optical flow to enhance temporal modeling, and motion-aware masking [9] highlights dynamic regions. SIGMA [33] replaces pixel-level targets with Sinkhorn-regularized cluster assignments, encouraging learning of semantic features. Yet, these approaches often rely on frozen teachers, handcrafted objectives, or sensitive clustering parameters. DISCOVR addresses these issues by introducing video self-distillation with evolving semantic guidance from an image encoder, aligning fine-grained spatial and temporal features to produce coherent, high-level video representations, without external supervision, handcrafted tasks, or modality-specific assumptions.

Self-Supervised Pretraining for Medical Videos. Given the limited availability of annotated data, several works have adapted video SSL techniques to medical domains. Jiao et al. [15] explored

https://doi.org/10.52202/085713-2045

frame order and transformation prediction for fetal ultrasound. EchoFlow [31] generated synthetic echocardiograms via adversarial VAEs and latent flow. Although effective in context, these methods inherit key limitations from natural video SSL, including reliance on frozen teachers, hand-crafted objectives, and sensitive clustering parameters. In addition, they adopt design choices tailored to natural images, such as short clip lengths and the lack of mechanisms for capturing fine-grained spatial cues, both of which are inadequate for clinical video analysis, where longer temporal context and detailed spatial reasoning are critical. In contrast, DISCOVR uses long (64-frame) clips and introduces dynamic semantic guidance from an evolving image encoder, enabling the video backbone to learn rich, fine-grained spatio-temporal representations without reliance on pretrained models or handcrafted supervision.

3 Methodology

The modelling of echocardiography video-based tasks poses unique challenges, as models must simultaneously detect fine-grained anatomical details, such as subtle septal defects, and accurately track how these features evolve throughout the cardiac cycle to reliably identify anomalies. We propose a unified self-supervised framework addressing these aspects without relying on labelled data or external pretrained models. Our method integrates three complementary techniques: (1) video self-distillation to capture global cardiac motion, (2) online spatial guidance to learn fine-grained structural information, and (3) semantic cluster distillation (SCD) loss to transfer fine-grained semantic knowledge from the evolving image encoder to the video model.

3.1 Video Self-Distillation

To capture how cardiac structures evolve throughout the cardiac cycle, it is essential to learn spatio-temporal representations from echocardiography videos. We propose a video-level self-distillation framework based on a student-teacher architecture with Vision Transformer (ViT)-based encoders (Fig. 2) that models temporal dynamics and improves understanding of global heart motion. Given a video input v, we partition it into non-overlapping 3D space-time patches (tube tokens), and prepend a learnable class (CLS) token, resulting in a sequence $x_0, x_1, \ldots, x_N$, where x_0 is the CLS token.

The teacher encoder E_{θ_t} processes the complete, unmasked video to produce a global representation, whereas the student encoder E_{θ_s} processes multiple masked variants $v_{\mathcal{M}1}, \ldots, v_{\mathcal{M}_M}$, each applying distinct random space-time masks to enforce inference of missing content.

Both encoders output a global video representation via the CLS token:

$$z_t = E_{\theta_t}(v)[0], \quad z_s^{(m)} = E_{\theta_s}(v_{\mathcal{M}_m})[0]. \tag{1}$$

The teacher parameters are updated using an exponential moving average (EMA) of the student parameters:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s, \quad \lambda \in [0,1). \tag{2}$$

These CLS embeddings are subsequently mapped through linear projection heads characterized by learnable weight matrices W_t (teacher) and W_s (student). The resulting embeddings are transformed into probability distributions via temperature-scaled softmax operations:

$$P_t = \text{softmax}\left(\frac{W_t z_t}{\tau_t}\right), \quad P_s^{(m)} = \text{softmax}\left(\frac{W_s z_s^{(m)}}{\tau_s}\right), \tag{3}$$

where τ_t and τ_s are temperature parameters for the teacher and student, respectively.

We align these probability distributions using the cross-entropy loss:

$$\mathcal{L}_{\text{ssl}}^{vid} = \frac{1}{M}\sum_{m=1}^{M} H(P_t, P_s^{(m)}), \tag{4}$$

where H denotes cross-entropy. This approach encourages the student to match the teacher's global representation of cardiac motion, despite observing only incomplete views of the video. Through video-level self-distillation, the student learns to recover the evolving dynamics of anatomical landmarks, capturing coherent motion patterns and structural features relevant to global heart function throughout the cardiac cycle.

3.2 Fine-Grained Online Spatial Guidance

Although video self-distillation promotes temporal consistency and global abstraction, it tends to overlook fine-grained spatial features, particularly those critical to clinical interpretation in echocardiography. Echocardiography imaging captures the dynamics and appearance of anatomically complex structures, where capturing subtle spatial details, such as mitral valve leaflet motion, septal wall thickness, or endocardial border definition, is crucial. To address this, we introduce a two-part strategy for enriching spatial detail and semantic structure in video representations:

a). Masked Image Self-Distillation. An online image encoder is trained to learn spatially rich features from partially masked images, enabling the extraction of fine-grained semantic concepts.
b). Semantic Cluster Distillation (SCD). A cross-modal clustering objective aligns reconstructed video tokens with spatial image features, encouraging the video model to organize its representation space around semantically meaningful structures.

3.2.1 Masked Image Self-Distillation

To learn fine-grained semantic features, we train an image encoder $\mathcal{I}_\theta$ in parallel with the video encoder. Each video v is decomposed into individual frames $\{x_t\}$, which are processed independently. For each frame x, the teacher image encoder $\mathcal{I}_{\theta_t}$ receives the full-resolution image, while the student encoder $\mathcal{I}_{\theta_s}$ is given N randomly masked variants $\{x_{\mathcal{M}_i}\}_{i=1}^N$. Each output is projected using distinct learnable heads W_t (teacher) and W_s (student), followed by softmax normalization:

$$P_s^{(i)} = \text{softmax}\left(\frac{W_s \mathcal{I}_{\theta_s}(x_{\mathcal{M}_i})}{\tau_s}\right), \quad P_t = \text{softmax}\left(\frac{W_t \mathcal{I}_{\theta_t}(x)}{\tau_t}\right), \tag{5}$$

where τ_s and τ_t are temperature parameters. The loss function encourages the student to match the teacher's predictions across all masked views:

$$\mathcal{L}_{\text{ssl}}^{img} = \frac{1}{N}\sum_{i=1}^{N} H(P_t, P_s^{(i)}), \tag{6}$$

with $H(\cdot,\cdot)$ denoting the cross-entropy. This training objective promotes the emergence of spatially grounded representations that encode fine-grained clinical concepts such as fetal heart valves, ventricular anatomy, and septal delineation that may be underrepresented in purely temporal learning.

3.2.2 Semantic Cluster Distillation (SCD)

While Masked Image Self-Distillation enables the image encoder to learn spatially grounded representations that capture fine-grained clinical concepts, it does not transfer this knowledge to the video encoder. As a result, the spatial and temporal representations remain disjoint. To bridge this gap, we introduce *Semantic Cluster Distillation (SCD)*, a cross-modal objective that distills semantic structure from the image encoder, guiding the video encoder to incorporate fine-grained spatial detail into its token representations.

Given a masked video input, the student video encoder E_{θ_s} processes the visible tokens to produce latent representations, which are then passed to a decoder ψ that reconstructs token-level features $\hat{\mathbf{z}}_v \in \mathbb{R}^{B\times N\times D}$, where B is the batch size, N is the number of masked tokens, and D is the feature dimension. In parallel, the corresponding video frames are processed by the image encoder $\mathcal{I}_{\theta_t}$, producing spatial features $\hat{\mathbf{z}}_i \in \mathbb{R}^{B\times N\times D}$. These image features are detached from the gradient flow and serve as semantic targets. Both sets of features are projected onto a shared set of learnable prototypes $P \in \mathbb{R}^{K\times D}$, resulting in similarity scores:

$$\mathbf{s}_v = \frac{\hat{\mathbf{z}}_v P^\top}{\tau}, \quad \mathbf{s}_i = \frac{\hat{\mathbf{z}}_i P^\top}{\tau}, \tag{7}$$

where τ is a temperature scaling parameter and K is the number of prototypes. The resulting scores are transformed into Sinkhorn soft cluster targets using the Sinkhorn-Knopp algorithm:

$$\mathbf{q}_v = \text{Sinkhorn}(\mathbf{s}_v), \quad \mathbf{q}_i = \text{Sinkhorn}(\mathbf{s}_i). \tag{8}$$

The SCD loss symmetrically aligns the two modalities by minimizing the cross-entropy between their soft cluster assignments:

$$\mathcal{L}_{\text{SCD}} = \text{CE}(\mathbf{s}_v, \text{stopgrad}(\mathbf{q}_i)) + \text{CE}(\mathbf{s}_i, \text{stopgrad}(\mathbf{q}_v)), \tag{9}$$

https://doi.org/10.52202/085713-2045

Table 1: **Comparison of video anomaly detection methods on three echocardiography datasets**. Our method consistently outperforms SOTA approaches, demonstrating improved effectiveness in identifying cardiac abnormalities across diverse patient populations.

Dataset	Model	Balanced Acc.	F1	AUC
EchoNet-Dynamic	MNAD	52.25	52.08	53.15
	MemAE	49.22	46.33	49.69
	C2FPL	57.36	57.35	59.00
	Ours	**63.20**	**61.45**	**67.06**
RVENET	MNAD	52.34	52.18	54.05
	MemAE	47.65	32.10	44.68
	C2FPL	47.88	47.86	46.30
	Ours	**56.23**	**53.88**	**57.42**
Echo Pediatric-LVH	MNAD	47.86	47.85	47.31
	MemAE	47.28	47.28	47.23
	C2FPL	51.39	51.31	50.68
	Ours	**55.63**	**54.63**	**57.23**

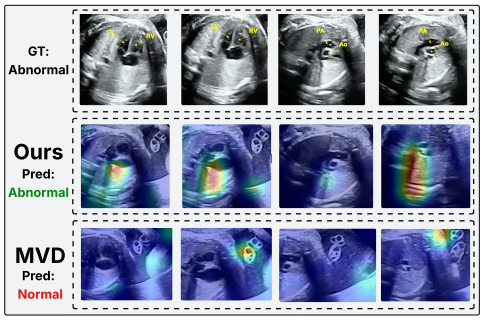

Figure 3: **Zero-Shot classification comparison:** (**Top**) The sweep from four-chamber to three-vessel view reveals smaller left-sided structures (LV and Ao) versus right-sided (RV and PA), consistent with coarctation of the aorta. (**Middle**) DISCOVR correctly identifies the abnormality, focusing on the ventricles in the four-chamber view and the Ao and PA in the vessel view. (**Bottom**)A backbone pretrained with MVD, in contrast, misclassifies the video as normal.

where gradients are propagated only through the video model and the prototype matrix P, while the image encoder is updated solely via its own self-distillation loss. This guides the video encoder to anchor its token representations to the spatially grounded clusters discovered by the image encoder, thereby distilling fine-grained anatomical detail into its temporal feature space. Semantic Cluster Distillation thus embeds spatial semantics within temporal features, yielding spatio-temporal representations that capture anatomically relevant detail in echocardiography videos.

4 Experiments and Results

Datasets. We use five ultrasound video datasets across fetal, pediatric, and adult populations. Two private fetal heart datasets, FetalEcho1 and FetalEcho2, were each collected from different hospital partners in the UK, comprising 10-second transverse, cephalad sweeps capturing five standard cardiac views (Situs, 4CH, LVOT, 3VV, 3VT). FetalEcho1 includes 8273/414/317 and FetalEcho2 includes 4154/320/305 videos for training/validation/testing. For adult and pediatric echocardiography, we use 3 public datasets: EchoNet Dynamic (apical 4CH adult; 7378/1326/1326) [28], EchoPediatric LVH (parasternal long-axis pediatric; 7837/1592/1592) [8], and RVENet (right ventricular pediatric/adult; 2516/487/573) [20]. Videos for adult and pediatric populations are labeled as *normal* or *abnormal* based on ejection fraction (EF), with *abnormal* defined as EF < 45% or EF > 75% [7]. Fetal videos are labeled as *normal* or *abnormal* based on expert evaluation by two fetal cardiologists(+10 years of experience). For the downstream segmentation task, we utilize the CAMUS [19] dataset.

Evaluation. All baseline models use official implementations, with videos sampled in 64-frame clips at a stride of 3. We adopt space-time tube embeddings from VideoMAE [35], treating each $2 \times 16 \times 16$ cube as a token with 90% masking ratio. All models use a ViT base backbone with consistent configurations. We evaluate representations using **zero-shot classification** and **linear probing**. Zero-shot evaluation uses a weighted kNN classifier [42, 4] on frozen features, with k selected based on validation balanced accuracy. Linear probing trains a linear classifier for 30 epochs on a frozen backbone using a labeled validation set. During inference, each test video is divided into 64-frame clips and classified independently; a video is labeled abnormal if any clip is predicted abnormal. For segmentation evaluation, we add a linear layer followed by Conv2D upsampling blocks to generate pixel-level masks while keeping the backbone frozen.

Baselines. We compare DISCOVR with SOTA video SSL methods SIGMA [33], MGMAE [14], MVD [39], VideoMAE [35], and RAD-DINO [30], covering masked modeling, clustering, and dense feature learning. For anomaly detection, we include SOTA methods MNAD [29], MemAE [12],

Table 2: **Linear probing classification results on five echocardiography datasets** spanning fetal, adult, and pediatric populations. Our method achieves SOTA results, outperforming prior video SSL baselines and generalizing effectively across diverse clinical cohorts.

Dataset	Model	Acc	Bal. Acc.	F1
Fetal-Echo 1	VideoMAE	60.19	60.01	59.82
	MGMAE	59.55	59.40	59.30
	SIGMA	63.11	62.93	62.78
	Ours	**65.70**	**65.52**	**65.39**
Fetal-Echo 2	VideoMAE	56.39	53.12	51.60
	MGMAE	60.98	60.49	60.43
	SIGMA	56.07	56.06	55.81
	Ours	**65.25**	**63.53**	**63.59**
Echonet-Dynamic	VideoMAE	71.04	70.86	70.85
	MGMAE	61.84	61.81	61.81
	SIGMA	75.57	75.48	75.50
	Ours	**77.68**	**77.61**	**77.63**
Echo Pediatric-LVH	VideoMAE	60.87	60.94	60.71
	MGMAE	54.71	51.70	49.46
	SIGMA	58.42	57.27	57.24
	Ours	**62.81**	**61.64**	**61.66**
RVENET	VideoMAE	60.03	60.31	59.70
	MGMAE	59.16	59.15	59.15
	SIGMA	59.51	59.25	58.98
	Ours	**62.65**	**62.68**	**62.65**

Table 3: **Zero-shot evaluation across five echocardiography datasets** covering fetal, adult, and pediatric populations. Our method consistently outperforms existing video SSL baselines, demonstrating robust generalization across diverse clinical populations.

Dataset	Population	Model	Acc	Bal. Acc.	F1
Fetal-Echo 1	Fetal	RAD-DINO	55.34	55.35	55.34
		VideoMAE	60.52	60.81	60.00
		SIGMA	54.37	54.91	51.90
		MGMAE	60.84	61.03	60.64
		MVD	59.87	60.20	59.15
		Ours	**62.46**	**62.79**	**61.79**
Fetal-Echo 2	Fetal	RAD-DINO	54.10	51.46	50.62
		VideoMAE	50.49	48.01	47.21
		SIGMA	55.41	51.90	49.92
		MGMAE	59.34	56.71	56.09
		MVD	59.34	55.45	53.14
		Ours	**59.67**	**57.18**	**56.69**
Echonet-Dynamic	Adult	RAD-DINO	59.43	59.63	59.34
		VideoMAE	57.16	57.91	55.07
		SIGMA	53.47	54.46	49.04
		MGMAE	51.21	52.23	46.13
		MVD	60.11	60.94	57.56
		Ours	**62.59**	**63.20**	**61.45**
Echo Pediatric-LVH	Pediatric	RAD-DINO	53.14	52.27	52.26
		VideoMAE	51.57	53.98	50.47
		SIGMA	47.55	49.56	46.80
		MGMAE	46.61	48.91	45.45
		MVD	49.56	51.91	48.46
		Ours	**54.65**	**55.63**	**54.63**
RVENET	Adult, Pediatric	RAD-DINO	55.67	55.65	**55.65**
		VideoMAE	54.97	55.64	52.24
		SIGMA	52.36	53.18	47.64
		MGMAE	53.23	54.08	48.17
		MVD	54.62	55.12	53.17
		Ours	**55.67**	**56.23**	53.88

and C2FPL [2], which rely solely on spatial-temporal learning without external modules like object detectors, pose estimators, or optical flow, often tailored to natural images.

4.1 Comparison with Video Anomaly Detection Methods

Table 1 compares the anomaly detection performance of DISCOVR with several state-of-the-art approaches. DISCOVR achieves the highest F1 score for all datasets (61.45% for EchoNet Dynamic, 53.88% for RVENET, and 54.63% for EchoPediatric LVH) as well as the highest balanced accuracy (63.20%, 56.23%, and 55.63%, respectively), substantially outperforming C2FPL, MemAE, and MNAD for all reported metrics. C2FPL relies on a multi-stage pseudo-labeling process to enhance anomaly discrimination, while both MemAE and MNAD incorporate sophisticated memory mechanisms and feature regularization in their inference pipelines. These methods employ targeted, anomaly-specific inference strategies and complex architectures.
In contrast, DISCOVR builds on a simple self-supervised learning framework that jointly learns spatial and temporal features, utilizing only a straightforward zero shot kNN classifier at inference. DISCOVR not only achieves state-of-the-art scores, including the highest AUCs of 67.06 on EchoNet Dynamic, 57.42 on RVENET, and 57.23 on EchoPediatric LVH, but also demonstrates that richer spatio-temporal representations learned via simple SSL can offer more effective and efficient anomaly detection than more sophisticaed anomaly detection techniques without reliance on specialized or resource intensive modules.

Linear Probing. Table 2 shows that DISCOVR achieves the highest balanced accuracy and F1 score in linear probing for anomaly detection across all echocardiography datasets. For example, on Echonet Dynamic, DISCOVR attains an F1 of 77.63 compared to 75.50 for SIGMA, and on FetalEcho 2, achieves 63.59 versus 60.43 for MGMAE. These improvements are consistent across fetal, pediatric, and adult cohorts. While VideoMAE and MGMAE rely on high masking ratios and pixel-level reconstruction, their representations often miss subtle anatomical landmarks and temporally distributed abnormalities, reflecting a lack of deeper semantic abstraction. Clustering-based approaches such as SIGMA can capture some temporal variation but lack explicit semantic guidance, limiting their ability to identify clinically relevant landmarks. In contrast, DISCOVR leverages semantic supervision from the image encoder through online distillation, combined with

https://doi.org/10.52202/085713-2045

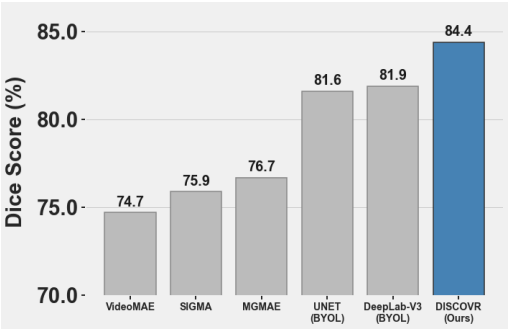

Figure 4: **Barplot comparing the segmentation performance across different models**. Our proposed DISCOVR approach achieves the highest Dice score of 0.844, outperforming both specialized segmentation architectures (DeepLab-V3, UNET) and other self-supervised methods.

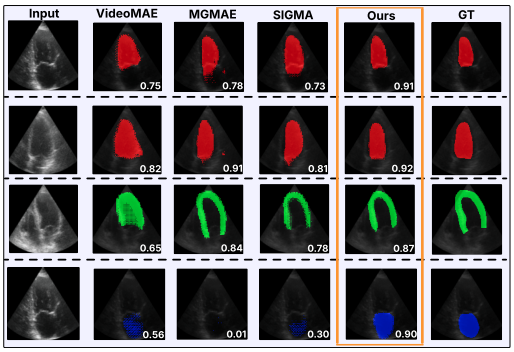

Figure 5: **Segmentation comparison on the CAMUS dataset** for left ventricular endocardium (LV Endo), left ventricular epicardium (LV Epi), and left atrium (LA). Our method produces accurate and consistent masks, achieving higher Dice scores (bottom right) than baseline methods.

temporal modeling in the video branch. This enables DISCOVR to capture fine-grained spatial features and their evolution over time, resulting in representations that are both robust and clinically meaningful for anomaly detection in cardiac ultrasound.

Zero-Shot. Table 3 shows that DISCOVR achieves the highest balanced accuracy and F1 score for zero shot classification across all echocardiography datasets. For example, on Echonet Dynamic, DISCOVR reaches an F1 of 61.45 compared to 57.56 for the best baseline, and on FetalEcho 1, achieves 61.79 versus 60.64 for MGMAE. These improvements are consistent across fetal, pediatric, and adult cardiac cohorts. This stronger performance reflects DISCOVR's ability to integrate semantic features captured by the image encoder with temporal dynamics modeled by the video branch, explicitly aligned through the SCD loss during self-supervised training. Pixel reconstruction models such as VideoMAE and MGMAE focus primarily on low-level appearance and texture, and clustering approaches like SIGMA, while using temporal clips, lack explicit semantic guidance. Image-based baselines like RAD-DINO do not leverage temporal information, while methods such as MVD that rely on external pretrained teachers may be less adaptable to the clinical and domain-specific challenges of ultrasound video. DISCOVR's capabilities are further highlighted in the qualitative example of Fig. 3, where it detects subtle cardiac structures and correctly classifies a challenging fetal video as abnormal, while MVD fails to capture these cues and predicts a normal outcome. This underscores how DISCOVR's features are sufficiently fine-grained to enable accurate zero shot anomaly detection, even without task-specific tuning.

4.2 Segmentation Evaluation

We evaluate the effectiveness of DISCOVR representations for downstream cardiac segmentation using the CAMUS dataset [19]. As shown in Fig 4, DISCOVR achieves the highest Dice score (0.844), outperforming specialized segmentation architectures such as UNet and DeepLabV3 (0.816 and 0.819, respectively, both with BYOL pretraining).When compared using a simple linear+upsampling head on a frozen backbone, DISCOVR also surpasses other SSL-based video models, including VideoMAE (0.747), MGMAE (0.767), and SIGMA (0.759). Fig. 5 highlights these advantages: DISCOVR produces consistently accurate and well-aligned segmentation masks for LV Endo, LV Epi, and especially the left atrium. For the challenging left atrium segmentation (blue mask), MGMAE misses the structure entirely (Dice = 0.01), while SIGMA and VideoMAE also perform poorly (Dice = 0.30 and 0.56). DISCOVR, in comparison, achieves 0.90, demonstrating superior ability to segment subtle structures and delineate boundaries due to its fine-grained feature learning.

4.3 LVEF Prediction

We evaluate the effectiveness of DISCOVR representations for downstream cardiac function estimation using the EchoNet-Dynamic ejection fraction dataset [28]. As shown in Table 4, DISCOVR achieves the lowest Mean Absolute Error (MAE) of 7.79 under the standard linear probing setup, outperforming other self-supervised baselines such as VideoMAE (8.02) and MGMAE (8.88). When

fine-tuning only the last three encoder blocks, DISCOVR further reduces the MAE to 6.32, demonstrating the strength of its learned representations even with limited adaptation. In comparison, the fully supervised EchoNet-Dynamic model [28] is trained end to end with all parameters updated. Under an ejection fraction–only setup without segmentation labels, DISCOVR surpasses these fully supervised baselines, including MC3 with an MAE of 6.59, the base EchoNet-Dynamic model with 7.35, and R3D with 7.63. The full EchoNet-Dynamic architecture achieves an MAE of 4.05 using a large multi-task design with 71.1 million parameters co-trained on 20,060 manual segmentation tracings. These results show that DISCOVR, through self-supervised pretraining and partial fine-tuning, learns powerful cardiac representations that rival or exceed fully supervised models trained end to end.

Table 4: LVEF prediction results on the EchoNet-Dynamic dataset. Our self-supervised method is compared against other SSL methods and fully-supervised baselines from [28].

Model	**MAE ↓**	**RMSE ↓**	**EF Labels**	**Seg. Labels**
Self-Supervised (Linear Probing)				
VideoMAE	8.02	11.16	✓	
MGMAE	8.88	12.47	✓	
DISCOVR (Ours)	**7.79**	**10.89**	✓	
Self-Supervised (Fine-tuning)				
DISCOVR (finetune last 3 blocks)	**6.32**	**8.62**	✓	
Fully-Supervised Baselines [1] trained only with EF Data				
MC3 (All frames)	6.59	9.39	✓	
EchoNet-Dynamic (EF, All frames)	7.35	9.53	✓	
R3D (All frames)	7.63	9.75	✓	
DISCOVR (finetune last 3 blocks,64 frames)	**6.32**	**8.62**	✓	
EchoNet-Dynamic (Full model)	**4.05**	**5.30**	✓	✓

5 Ablation Study

In this section, we ablate the key components of the training objective in our model, **DISCOVR**. All experiments are conducted on the **Echonet Dynamic** dataset and evaluated using the **k-nearest neighbor (kNN) protocol**. This setup allows us to assess the discriminative quality of the learned representations in a fully frozen setting without additional fine-tuning.

Effect of Loss Components. We evaluate the effect of two core loss components used in DISCOVR: (i) the video self-distillation component ($\mathcal{L}_{\text{ssl}}^{\text{vid}}$), and (ii) the semantic cluster distillation component with online image guidance ($\mathcal{L}_{SCD}$). Table 5a reports the performance of these losses individually and in combination in zero-shot settings. Using only $\mathcal{L}_{\text{ssl}}^{\text{vid}}$ yields modest performance (F1 = 48.23%), as it primarily captures global temporal structure via CLS tokens but lacks guidance for fine-grained semantics. Introducing $\mathcal{L}_{SCD}$ leads to a substantial improvement (F1 = 61.45%, Balanced Accuracy = 63.20%), as the evolving image-based semantic clusters enrich the temporal features learned by the video model and encourage focus on more fine-grained, spatially grounded information. For more detailed ablation, refer to supplementary section B.1.4.

Effect of Backbone Size. We investigate how transformer backbone size impacts DISCOVR's representation quality. We evaluate ViT-Small and ViT-Base variants, each paired with matching DINO image encoders, on the Echonet Dynamic dataset using kNN evaluation (Table 5b). ViT-Base achieves superior performance (F1=61.45%, balanced accuracy=63.20%) compared to ViT-Small (F1=57.52%, balanced accuracy=59.44%). The smaller model's reasonable performance indicates DISCOVR learns meaningful representations even with limited capacity.

Effect of Number of Frames. In this ablation, we evaluate how the number of frames sampled from each video clip affects the representational quality learned by our model. We experiment with three temporal lengths: 16, 32, and 64 frames. All other training settings are kept constant, and the results are reported in Table 5d. We observe a clear upward trend in performance with increasing frame count. Using 16 frames results in an F1 score of 55.68%, which improves to 57.45% with 32 frames. The best performance is achieved with 64 frames, yielding an F1 score of 61.45% and balanced accuracy of 63.20%. These results support the intuition that ultrasound, being a temporally dense and

 https://doi.org/10.52202/085713-2045

Table 5: Ablation studies reported with Balanced Accuracy, Precision, and F1 score.

(a) Effect of loss terms

$\mathcal{L}_{\text{ssl}}^{vid}$	$\mathcal{L}_{SCD}$	Bal. Acc.	Precision	F1
✓	✗	52.27	53.16	48.23
✓	✓	**63.20**	**65.35**	**61.45**

(b) Backbone size

Backbone	Bal. Acc.	Precision	F1
ViT-Small	59.44	61.03	57.52
ViT-Base	**63.20**	**65.35**	**61.45**

(c) Masking ratio

Mask (%)	Bal. Acc.	Precision	F1
50	55.60	56.90	52.98
75	56.25	57.58	53.85
90	**63.20**	**65.35**	**61.45**

(d) Number of frames

Frames	Bal. Acc.	Precision	F1
16	57.89	59.45	55.68
32	59.54	61.36	57.45
64	**63.20**	**65.35**	**61.45**

dynamic modality, benefits from longer clips. More frames provide richer temporal context, enabling the model to capture fine-grained spatial and temporal motion patterns across the cardiac cycle.

Effect of Masking Ratio. Table 5c shows a steady improvement in performance as the masking ratio increases, with F1-score rising from 52.98% (50%) to 61.45% (90%). Higher masking forces both the video encoder and the semantic image guidance branch to infer more from sparse visual cues, encouraging the model to focus on the most salient and non-redundant features. This promotes the learning of richer representations that better capture subtle and fine-grained spatio-temporal patterns, resulting in improved anomaly detection performance.

Computational Cost and Scalability. We report the computational cost, of our method in Table 6. The table shows both training and inference statistics, showing GPU memory usage and F1-score for each method on EchoNet-Dynamic, for a batch size of 1, 64 frames, and a spatial size of 112×112. During training, DISCOVR uses slightly more GPU memory (10.5GB) compared to prior methods (between 9.0 and 9.5GB) but achieves a notable +6.38% improvement in F1-score over the closest competitor. At inference, all methods, including DISCOVR, use identical ViViT-like encoders, resulting in nearly the same GPU memory footprint and FLOPS. This demonstrates that our method's performance improvements come with minimal extra training cost and no penalty for inference efficiency.

Table 6: Training and inference GPU memory, FLOPS, and F1-score on EchoNet-Dynamic, batch size = 1, 16 frames, 112×112 resolution.

Model	Train Mem (GB)	F1-score	Infer Mem (GB)	Infer. FLOPS
MGMAE [14]	9.0	46.13	1.153	101.85
VideoMAE [35]	9.0	55.07	1.153	101.85
SIGMA [33]	9.2	49.04	1.153	101.85
Video-distillation	9.5	48.23	1.153	101.85
DISCOVR (Ours)	10.5	**61.45**	1.153	101.85

6 Conclusion

We introduce DISCOVR, a self-supervised model for learning video representations in echocardiography across diverse patient populations. Our approach combines masked video modeling, temporal self-distillation, and online spatial supervision, unified by a Semantic Cluster Distillation (SCD) objective that aligns video and image features through cross-modal clustering, without relying on labeled anomalies or pretrained models. Extensively evaluated on six echocardiography datasets spanning fetal, pediatric, and adult populations, DISCOVR consistently outperforms previous self-supervised and anomaly detection methods for multiple tasks, including anomaly detection, classification (zero-shot and linear probing), and segmentation. DISCOVR's task-agnostic design and its applicability to diverse patient groups establish it as a strong foundation for screening cardiac conditions and developing assistive tools for echocardiography.

Acknowledgments

We acknowledge financial support from InnoHK-funded Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE), UKRI grant EP/X040186/1, UK EPSRC grant EP/T028572/1 (VisualAI), UK EPSRC Doctoral Training Partnership award and, UKRI AIRR Early Access Project No. ANON-BYYG-VX4C-Z.

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

[2] Anas Al-Lahham, Nurbek Tastan, Muhammad Zaigham Zaheer, and Karthik Nandakumar. A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6793–6802, 2024.

[3] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9922–9931, 2020.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[6] Yutong Chen, Hongzuo Xu, Guansong Pang, Hezhe Qiao, Yuan Zhou, and Mingsheng Shang. Self-supervised spatial-temporal normality learning for time series anomaly detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–162. Springer, 2024.

[7] Cleveland Clinic. Ejection fraction: What it is, types and normal range, 2023. Accessed: 2025-05-01.

[8] 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.

[9] David Fan, Jue Wang, Shuai Liao, Yi Zhu, Vimal Bhat, Hector Santos-Villalobos, Rohith MV, and Xinyu Li. Motion-guided masking for spatiotemporal representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5619–5629, 2023.

[10] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

[11] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

[12] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714, 2019.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

 https://doi.org/10.52202/085713-2045

[14] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13493–13504, 2023.

[15] Jianbo Jiao, Richard Droste, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Self-supervised representation learning for ultrasound video. In *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*, pages 1847–1850. IEEE, 2020.

[16] Dahun Kim, Donghyeon Cho, and In So Kweon. Self-supervised video representation learning with space-time cubic puzzles. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8545–8552, 2019.

[17] RL Knowles and RM Hunter. Screening for congenital heart defects: external review against programme appraisal criteria for the uk nsc. 2014.

[18] RM Lang, LP Badano, V Mor-Avi, 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. *J. Am. Soc. Echocardiogr*, 28(1):1–39, 2015.

[19] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019.

[20] Bálint Magyar, Márton Tokodi, András Soós, Máté Tolvaj, Bálint Károly Lakatos, Alexandra Fábián, Elena Surkova, Béla Merkely, Attila Kovács, and András Horváth. Rvenet: a large echocardiographic dataset for the deep learning-based assessment of right ventricular function. In *European Conference on Computer Vision*, pages 569–583. Springer, 2022.

[21] Divyanshu Mishra, Pramit Saha, He Zhao, Netzahualcoyotl Hernandez-Cruz, Olga Patey, Aris Papageorghiou, and J Alison Noble. Mcat: Visual query-based localization of standard anatomical clips in fetal ultrasound videos using multi-tier class-aware token transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28267–28275, 2025.

[22] Divyanshu Mishra, Pramit Saha, He Zhao, Netzahualcoyotl Hernandez-Cruz, Olga Patey, Aris T Papageorghiou, and J Alison Noble. Tier-loc: Visual query-based video clip localization in fetal ultrasound videos with a multi-tier transformer. *Medical Image Analysis*, page 103611, 2025.

[23] Divyanshu Mishra, Pramit Saha, He Zhao, Olga Patey, Aris T Papageorghiou, and J Alison Noble. Stan-loc: Visual query-based video clip localization for fetal ultrasound sweep videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 742–752. Springer, 2024.

[24] Divyanshu Mishra, He Zhao, Pramit Saha, Aris T Papageorghiou, and J Alison Noble. Dual conditioned diffusion models for out-of-distribution detection: Application to fetal ultrasound videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 216–226. Springer, 2023.

[25] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 527–544. Springer, 2016.

[26] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*, 2018.

[28] 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.

[29] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14372–14381, 2020.

[30] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Rad-dino: Exploring scalable medical image encoders beyond text supervision. *arXiv preprint arXiv:2401.10815*, 2024.

[31] Hadrien Reynaud, Alberto Gomez, Paul Leeson, Qingjie Meng, and Bernhard Kainz. Echoflow: A foundation model for cardiac ultrasound image and video generation. *arXiv preprint arXiv:2503.22357*, 2025.

[32] Pramit Saha, Divyanshu Mishra, Netzahualcoyotl Hernandez-Cruz, Olga Patey, Aris Papageorghiou, Yuki M Asano, and J Alison Noble. Self-supervised normality learning and divergence vector-guided model merging for zero-shot congenital heart disease detection in fetal ultrasound videos. *arXiv preprint arXiv:2503.07799*, 2025.

[33] Mohammadreza Salehi, Michael Dorkenwald, Fida Mohammad Thoker, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. Sigma: Sinkhorn-guided masked video modeling. In *European Conference on Computer Vision*, pages 293–312. Springer, 2024.

[34] Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H Li, Mingkui Tan, and Chuang Gan. Masked motion encoding for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2235–2245, 2023.

[35] 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.

[36] Denise van der Linde, Estelle EM Konings, Mirjam A Slager, Maarten Witsenburg, Willem A Helbing, Johanna JM Takkenberg, and Jolien W Roos-Hesselink. Birth prevalence of congenital heart disease worldwide: a systematic review and meta-analysis. *Journal of the American College of Cardiology*, 58(21):2241–2247, 2011.

[37] Amber EL van Nisselrooij, AKK Teunissen, Sally-Ann Clur, Lieke Rozendaal, Eva Pajkrt, Ingeborg H Linskens, L Rammeloo, Jan MM van Lith, Nico A Blom, and Monique C Haak. Why are congenital heart defects being missed? *Ultrasound in Obstetrics & Gynecology*, 55(6):747–757, 2020.

[38] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 504–521. Springer, 2020.

[39] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6312–6322, 2023.

[40] World Health Organization. Cardiovascular diseases (CVDs): Fact Sheet. `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`, 2021. Accessed: October 2023.

[41] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu. Self-supervised sparse representation for video anomaly detection. In *European Conference on Computer Vision*, pages 729–745. Springer, 2022.

[42] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

 https://doi.org/10.52202/085713-2045

[43] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10334–10343, 2019.

[44] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[45] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.

Appendix

A Dataset Distribution

This section presents the dataset distributions for our five echocardiography video datasets: FetalEcho1 (Fig.6), FetalEcho2 (Fig.7), EchoNet-Dynamic (Fig.8), EchoNet-Pediatric (Fig.9), and RVENET (Fig.10). For each dataset, the bar chart displays the number of unique samples in the training, validation, and test sets. The accompanying pie charts illustrate the class distributions (Normal vs. Abnormal) within the validation and test sets.

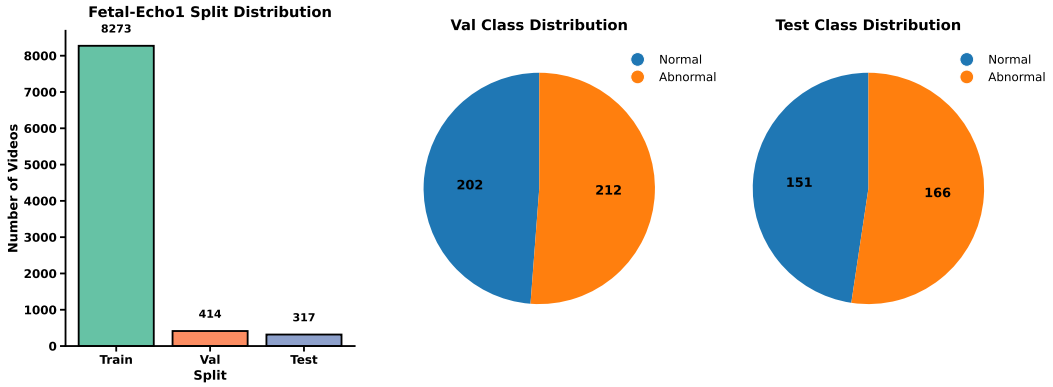

Figure 6: Dataset Distribution for Fetal-Echo1 dataset

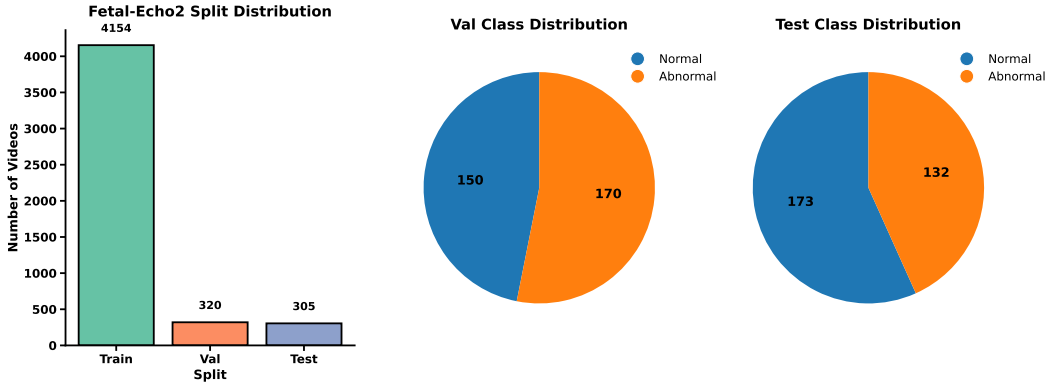

Figure 7: Dataset Distribution for Fetal-Echo2 dataset

https://doi.org/10.52202/085713-2045

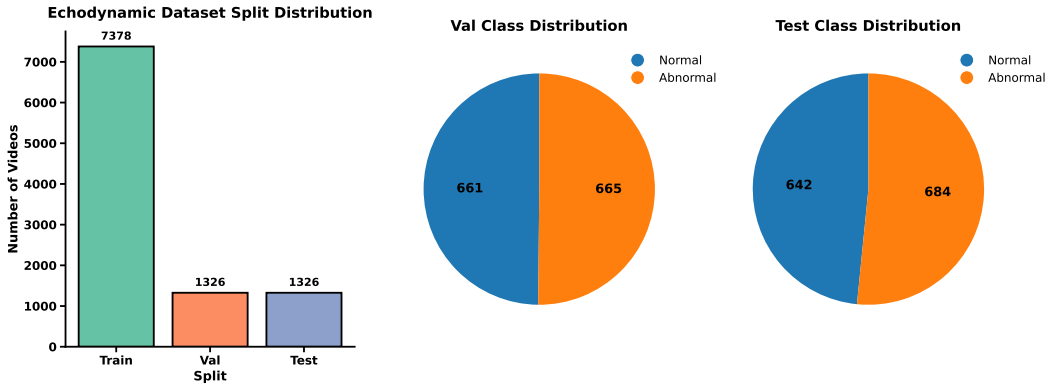

Figure 8: Dataset Distribution for Echo-Dynamic dataset

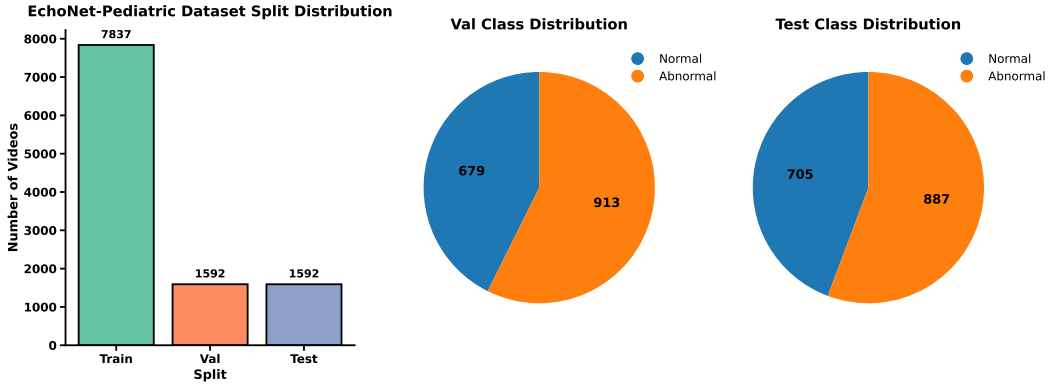

Figure 9: Dataset Distribution for Echo-Pediatric dataset

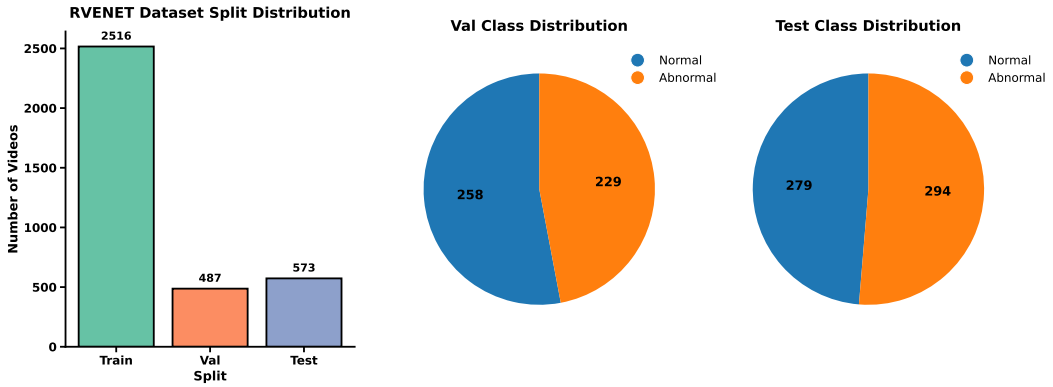

Figure 10: Dataset distribution for RVENET dataset

B Additional Results:

B.1 Full Finetuning

B.1.1 Evaluation Setup

We follow the same evaluation procedure as described in the experiments section in the main paper, but fine-tune the entire backbone along with the linear layer. All other evaluation settings remain

unchanged. Results are reported on the Echonet Dynamic dataset to assess end-to-end supervised performance.

B.1.2 Evaluation Result:

Under full fine-tuning, as shown in Table 7, all models experience a drop in performance compared to their linear probing results, reflecting overfitting due to the limited labeled validation data. Despite this, DISCOVR achieves the highest F1 score of 70.44%, outperforming MGMAE (65.99%), SIGMA (61.46%), and VideoMAE (57.31%). DISCOVR's structured representation learning, through temporal distillation and cross-modal clustering, appears to provide more robust and adaptable features, enabling it to generalize better even when fully fine-tuned on a small dataset.

Table 7: Table showing the full-finetuning result of DISCOVR compared to other baselines on the Echo-Dynamic Dataset

Model (%)	Accuracy	Balanced Acc.	Precision	Recall	F1-Score
VideoMAE	57.62	57.94	58.27	57.94	57.31
SIGMA	61.69	62.00	62.41	62.00	61.46
MGMAE	65.99	66.08	66.10	66.08	65.99
Ours	**70.51**	**70.42**	**70.50**	**70.42**	**70.44**

B.1.3 Generalisation to Other Modalities.

To test whether DISCOVR generalizes beyond echocardiography, we evaluated its transfer performance on two distinct medical image benchmarks: the Breast Ultrasound Images dataset [1] (cancer detection across 600 patients) and DermMNIST [44] (skin lesion classification). Both breast ultrasound and echocardiography require the detection of small, irregular regions of altered tissue, such as hypoechoic tumors in the breast or localized wall motion abnormalities in the heart, making the ability to identify subtle structural changes in one domain directly applicable to the other. Similarly, DermMNIST demands fine-grained visual discrimination between morphologically similar skin lesions. For both benchmarks, we froze the DISCOVR encoder and trained a linear classifier, comparing performance directly across methods.

As shown in Table 8, our method demonstrates strong generalization across both tasks. On the Breast Ultrasound dataset, DISCOVR improves balanced accuracy by 2.01% over VideoMAE, 19.83% over SIGMA, and 12.01% over MGMAE. For DermMNIST, DISCOVR achieves an accuracy of 71.68%, outperforming VideoMAE by 2.85%, SIGMA by 3.00%, and MGMAE by 3.85%. These results demonstrate strong generalization to multiple medical image analysis tasks beyond echocardiography. Further, to assess generalization to natural video data, we pretrained and evaluated all models on the Kinetics 400 action recognition benchmark, using a zero-shot protocol where KNN classification with $K = 20$ was applied to features using 64 frames from the frozen video backbone. As shown in Table 9, DISCOVR achieves the highest Top-1 accuracy at 22.3%, outperforming MVD by 3.6%, MME by 3.2%, and VideoMAE by 1.6%, while also requiring the fewest pretraining epochs. These results highlight that DISCOVR not only excels at medical video tasks but also learns generalizable representations efficiently for large-scale natural video datasets.

Table 8: Linear Probing results on the **Breast Ultrasound** dataset and **DermMNIST**. For Breast Ultrasound, we report Balanced Accuracy (Bal. Acc.) and F1; for DermMNIST, we report overall Accuracy (Acc.).

Method	Breast Ultrasound		DermMNIST
	Balanced Accuracy	F1 Score	Accuracy
VideoMAE	61.75	64.45	68.83
SIGMA	43.93	42.21	68.68
MGMAE	51.75	52.34	67.83
DISCOVR (Ours)	**63.76**	**65.44**	**71.68**

https://doi.org/10.52202/085713-2045

Table 9: Zero-Shot KNN classification performance on Kinetics-400

Model	Epochs	Top-1 Accuracy (%)
MVD [39]	1600	18.7
MME [34]	800	19.1
VideoMAE [35]	800	20.7
DISCOVR (Ours)	**400**	**22.30**

B.1.4 Loss function Ablation detailed.

To rigorously evaluate each component, we have added baselines using only masked image or only video self-distillation. Indeed, we find that the settings perform suboptimally, as shown in Table 1, confirming that spatial or temporal cues alone are insufficient for strong representation learning. In contrast, combining both with the SCD loss, which explicitly distills fine-grained semantic structure from the image branch into the video backbone, achieves the best results. This supports our intuition that SCD is crucial for aligning spatial semantics with temporal dynamics, enabling more robust and clinically meaningful video representations.

Table 10: Effect of the different loss terms on classification performance (Balanced Accuracy, Precision, and F1).

$\mathcal{L}_{\text{ssl}}^{vid}$	$\mathcal{L}_{\text{ssl}}^{img}$	$\mathcal{L}_{\text{SCD}}$	Bal. Acc.	Precision	F1
✓	✗	✗	52.27	53.16	48.23
✗	✓	✗	53.66	55.22	49.43
✓	✓	✓	**63.20**	**65.35**	**61.45**

C Implementation Details

All models are implemented in PyTorch 2.6 and trained on RTX 8000 GPUs (48 GB) with a batch size of 8 using the AdamW optimizer. Videos are processed as 64-frame clips sampled at a stride of 3 and resized to 112×112.

For both video and image self-distillation, we use a student-teacher setup where the teacher processes the full input and the student observes $N = 4$ randomly masked views. The teacher network is updated via an exponential moving average (EMA) of the student with momentum $\lambda = 0.996$. A fixed temperature $\tau_s = 0.1$ is used for the student, while the teacher temperature τ_t is linearly warmed from 0.04 to 0.07 over the first 30 epochs. Semantic Cluster Distillation (SCD) uses $K = 3000$ learnable prototypes, with similarity scores computed via temperature-scaled dot products ($\tau = 0.1$) and cluster assignments generated using the Sinkhorn-Knopp algorithm (10 iterations, $\epsilon = 0.05$). Models are trained for 400 epochs with a learning rate of 1.5×10^{-4}, weight decay of 0.05, and 40 warmup epochs.

D Broader Impact and Limitations

In this work, we introduce **DISCOVR**, a novel self-supervised model for echocardiography video understanding across fetal, pediatric, and adult populations. Trained without labeled abnormal cases, DISCOVR learns rich spatiotemporal representations and enables zero-shot inference. One key application of DISCOVR is in the *early screening of heart diseases*, where it can assist clinicians by flagging potential anomalies in echocardiography videos. This has significant clinical relevance, as congenital heart defects affect approximately 1 in 100 newborns, with up to 50% missed during prenatal screening [36, 17, 37], and cardiovascular diseases remain the leading global cause of death [40]. By reducing reliance on large, labeled datasets, DISCOVR offers a scalable and accessible solution, particularly for deployment in low-resource settings.

While DISCOVR shows strong potential, its current scope is focused specifically on echocardiography, and it has not yet been evaluated on other imaging modalities. The model was trained and tested on

five datasets collected from distinct clinical sites, each with its own imaging protocols, devices, and patient cohorts. As a result, the demographic and geographic diversity of the data may be limited. Further validation is needed to assess the model's generalizability across broader clinical settings, populations, and imaging systems.

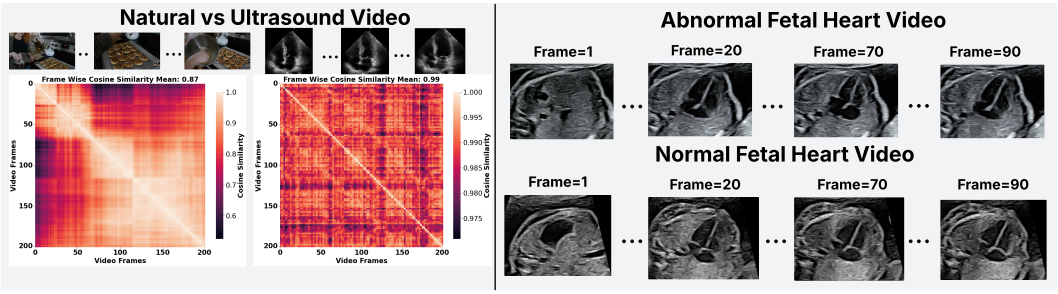

Figure 11: Figure (left) compares two fine-grained videos: a natural scene of a person baking (left) and an adult heart ultrasound (right). The frame-level cosine similarity matrix, computed using a pretrained VideoMAE model, shows that ultrasound frames are highly similar (mean=0.99), with only minor local variations. This highlights the difficulty in distinguishing individual frames in such medical videos. Figure (right) compares normal and abnormal fetal echocardiograms, which appear almost identical despite one being abnormal. This illustrates the inherent difficulty of distinguishing subtle cardiac abnormalities in fetal imaging.

https://doi.org/10.52202/085713-2045

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In this paper, we develop a self-supervised video foundational model for echocardiography and show its performance across fetal, pediatric and adult population across various downstream tasks. Our novelty lies in the DISCOVR framework, which combines a clustering-based video encoder for temporal dynamics with a DINO-inspired image encoder for spatial semantics, unified through a novel online semantic distillation loss that transfers anatomical knowledge from the evolving image encoder to the video encoder without relying on labels or external pretrained models.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our model is focused on echocardiography, and we have made it clear in the introduction/abstract and conclusion. Further, we have added a broader impact and limitations section in the appendix to discuss our assumptions and limitations of the current work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: In our paper, we have no theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have a dedicated Experiments section where we detail all implementation specifications, evaluation protocols, and dataset characteristics needed for reproducing our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

 https://doi.org/10.52202/085713-2045

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release the code and model weights upon acceptance. The paper utilizes four open-source datasets. The additional two fetal datasets are private, and we can not release them publicly due to data governance requirements of our study/hospital partner agreements. We have described the data characteristics for these datasets.

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have a dedicated section detailing experimental settings, evaluation, and dataset details. We have also included the dataset split visualization and detailed implementation details in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: Reporting classical confidence intervals or p-values would require training each configuration multiple times to estimate variance, an approach that is impractical for video self-supervised learning (SSL) methods, which typically demand substantial computational resources and incur high environmental costs. Instead, we assess robustness by evaluating our model architecture on five distinct echocardiography datasets covering fetal, pediatric, and adult populations. Across all settings, our method consistently outperforms strong baselines by a large margin, demonstrating reliable generalization without requiring repeated runs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information is mentioned in the Experiment and Results section in main paper and the implementation section of the Appendix.

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the work is built using anonymized patient data and conforms with all other NeurIPS guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

https://doi.org/10.52202/085713-2045

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We mention the societal impact of our work which has potential to support experts in detecting heart diseases across diverse patient populations and enhance clinical workflows through various downstream applications in the abstract, introduction, and conclusion sections. We have also added a separate Broad Impact and Limitations section in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cited all sources whose data or code we have utilized. Descriptions of each dataset are provided below:

(a) **Fetal-Echo1 and Fetal-Echo2**: These are private datasets collected as part of a private project. Ethics approval has been obtained, and all data were anonymised before model development. We will provide the ethics approval number and any other required documentation upon publication, in compliance with the double-blind review policy.

(b) **EchoNet-Dynamic and EchoNet-Pediatric**: Both are publicly available datasets released by Stanford University for non-commercial research use under a Stanford University Research Use Agreement, which we adhere to. For more information, see: `https://echonet.github.io/dynamic/`, `https://echonet.github.io/pediatric/`.

(c) **RVENet**: This dataset is available for non-commercial research use under a Research Use Agreement with Semmelweis University, which we comply with. For more information, see: `https://rvenet.github.io/dataset/`.

(d) **CAMUS**: This dataset is publicly available for research use and requires citation of the original publication, which we have provided. For more information, see: `https://www.creatis.insa-lyon.fr/Challenge/camus/databases.html`.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We will be releasing code, model weights and will be mentioning the license of use in the GitHub page.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

https://doi.org/10.52202/085713-2045

Answer: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The paper utilizes two private datasets, FetalEcho1 and FetalEcho2, collected from our partner hospitals. We have obtained institutional approvals, including ethics clearance and data anonymization approval. Specific details, such as the ethics approval number, will be provided upon acceptance to comply with the double-blind review policy.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.