ECHO-VISION-FM: A PRE-TRAINING AND FINE-TUNING FRAMEWORK FOR ECHOCARDIOGRAM VIDEOS VISION FOUNDATION MODEL

Ziyang Zhang[1], **Qinxin Wu**[2], **Sirui Ding**[3], **Xiaolong Wang**[1], and **Jiancheng Ye** [4*]

[1]Department of Electrical and Computer Engineering, Northwestern University
[2]Polytechnic Institute, Zhejiang University
[3]Bakar Computational Health Sciences Institute, University of California San Francisco
[4]Weill Cornell Medicine, Cornell University

ABSTRACT

Background: Echocardiograms provide vital insights into cardiac health, but their complex, multi-dimensional data presents challenges for analysis and interpretation. Current deep learning models for echocardiogram analysis often rely on supervised training, limiting their generalizability and robustness across datasets and clinical environments.

Objective: To develop and evaluate **EchoVisionFM** (**E**chocardiogram video **Vision F**oundation **M**odel), a self-supervised video learning framework designed to pre-train a video encoder on large-scale, unlabeled echocardiogram data. EchoVisionFM aims to produce robust and transferrable spatiotemporal representations, improving downstream performance across diverse echocardiogram datasets and clinical conditions.

Methods: Our framework employs Echo-VideoMAE, an autoencoder-based video transformer that compresses and reconstructs echocardiogram video data by masking non-overlapping video patches and leveraging a ViT encoder-decoder structure. For enhanced representation, we introduce **STFF-Net**, a **S**patio **T**emporal **F**eature **F**usion **Net**work, to integrate spatial and temporal features from the manifold representations. We pre-trained EchoVisionFM using the MIMIC-IV-ECHO dataset and fine-tuned it on the EchoNet-Dynamic dataset for downstream tasks, including classification and regression of key cardiac parameters.

Results: EchoVisionFM demonstrated superior performance in classifying left ventricular ejection fraction (LVEF), achieving an accuracy of 89.12%, an F1 score of 0.9323, and an AUC of 0.9364. In regression tasks, EchoVisionFM outperformed state-of-the-art models, with LVEF prediction reaching a mean absolute error (MAE) of 4.18% and an R^2 of 0.8022. The model also showed significant improvements in estimating end-systolic and end-diastolic volumes, with R^2 values of 0.8006 and 0.7296, respectively. Incorporating STFF-Net led to further performance gains across tasks.

Conclusion: Our results indicate that large-scale self-supervised pre-training on echocardiogram videos enables the extraction of transferable and clinically relevant features, outperforming traditional CNN-based methods. The EchoVisionFM framework, particularly with STFF-Net, enhances the extraction of spatiotemporal features, improving the predictive accuracy for various cardiac parameters. EchoVisionFM offers a powerful, scalable approach for echocardiogram analysis, with potential applications in clinical diagnostics and research.

Keywords: Echocardiogram, Self-supervised video learning, Masked auto-encoder, Spatial-temporal feature merging, Foundation model, Cardiovascular disease diagnosis

*Corresponding author: Jiancheng Ye, jiancheng.ye@u.northwestern.edu

1 Introduction

Echocardiograms are among the most widely used non-invasive imaging techniques in cardiology, essential for assessing cardiac structure and function.[1] The dynamic nature of echocardiogram videos provides a rich dataset for diagnosing various heart conditions, including valvular heart disease, cardiomyopathies, heart failure, and congenital heart disease.[2] The clinical utility of echocardiogram lies in its ability to deliver real-time imaging of cardiac anatomy and function without exposing patients to radiation, making it a safer alternative to imaging modalities such as CT scans and X-rays. Echocardiograms offer invaluable insights into heart morphology, blood flow, and tissue motion, aiding clinicians in diagnosing conditions, planning treatment strategies, and monitoring therapeutic interventions.[1] Their diagnostic capabilities are diverse, enabling visualization of left ventricular size and function, assessment of valvular abnormalities, detection of pericardial effusions, and evaluation of myocardial ischemia.[2] Additionally, their rapid availability and cost-effectiveness enhance their appeal in clinical settings.

Despite their widespread use, interpreting echocardiograms poses challenges. Cardiologists and echocardiographers rely on manual interpretation, a process that demands significant expertise and is inherently subjective.[3] Variability in image quality, inter-operator differences, and the need for experience in distinguishing normal from pathological findings contribute to the complexity of echocardiogram interpretation. As global demand for echocardiogram increases, the shortage of expert clinicians becomes evident, particularly in under-resourced healthcare systems. Furthermore, the complexity of echocardiographic images, influenced by patient body composition and operator skill, complicates the standardization of interpretations and leads to inter-observer variability.[4]

In the past decade, advancements in deep learning and computer vision have led to the development of numerous vision models that impact various industries, including healthcare. Recently, the application of Vision Foundation Models (VFMs)[5, 6, 7, 8, 9, 10] in medical imaging analysis has shown great promise, particularly in echocardiogram.[11] These advanced machine learning frameworks, pre-trained on extensive datasets, provide a robust foundation for constructing specialized models that enhance the accuracy and efficiency of analyzing cardiac ultrasound images.[12, 3, 13, 14, 15] By leveraging VFMs, clinicians and researchers can utilize deep learning capabilities to automatically recognize and quantify cardiac structures and functions, potentially transforming the diagnostic landscape. This approach promises improved diagnostic accuracy and a significant reduction in variability associated with human interpretation, paving the way for more tailored and expedient cardiac treatment.

Traditionally, echocardiogram analysis has heavily relied on the interpretation of static two-dimensional frames, [16, 17, 18, 19, 20, 21, 22] which can limit understanding due to the dynamic nature of cardiac function. This reliance results in the loss of crucial temporal and three-dimensional spatial information necessary for assessing cardiac dynamics.[23, 24, 25] Static frames provide only snapshots, omitting continuous motion and volumetric changes of the heart throughout the cardiac cycle, which are vital for evaluating parameters such as ejection fraction, wall motion, and blood flow patterns. This dimensionality reduction can lead to inaccuracies and oversights in diagnoses, as critical aspects of heart function between frames may be missed. Additionally, dependence on single-frame interpretations heightens variability and subjectivity among different interpreters, potentially resulting in inconsistent clinical outcomes. The inability to capture and analyze the full spectrum of cardiac activity in real time poses significant challenges for accurate diagnosis and effective monitoring of heart conditions.

To address these limitations, we employed one of the most popular and robust video VFMs, VideoMAE (Video Masked Auto-Encoder), to directly model video data and generate a sequence of video manifold representations.[26] By utilizing VideoMAE, we gain a comprehensive view of the heart's movements and functionalities over time. This methodological shift allows us to extract temporal and spatial features that more accurately represent physiological conditions, providing deeper insights into cardiac mechanics. The capacity to model sequential video data facilitates continuous assessment of the heart, enabling the identification of subtle changes that might be overlooked when analyzing isolated frames. This approach enhances the precision and reliability of cardiac assessments and propels us toward more sophisticated, automated diagnostic tools that can improve patient outcomes through early and accurate detection of cardiac anomalies. We adapted the vanilla VideoMAE for the echocardiogram domain, training the model on a public dataset, resulting in our **Echo-VideoMAE**.

An important question arises regarding the optimal use of video representations generated by VideoMAE. In traditional 2D image processing, pooling all representations to create the most expressive representation is common, whether derived from ConvNets [27, 28, 29] or Transformers-style models.[30, 31] However, we argue that applying the same strategy to video data would squander the spatiotemporal correlations inherent in these representations. To improve cardiac function assessment and other heart-related tasks, we propose a concise, efficient fine-tuning network, called the **S**patio **T**emporal **F**eature **F**usion network (**STFF-Net**), to enhance the performance of echocardiogram diagnoses.

In this paper, we develop **EchoVisionFM** (**E**chocardiogram video **Vision F**oundation **M**odel), an effective two-stage deep learning framework for automating echocardiogram diagnosis. First, Echo-VideoMAE is trained on the MIMIC-

IV-ECHO dataset in a self-supervised manner. Next, STFF-Net is applied on top of the well-trained echo video encoder, fine-tuning all trainable parameters for specific downstream tasks. Our extensive experiments demonstrate that our proposed method surpasses state-of-the-art models in echocardiogram tasks. In summary, our contributions are as follows:

- **Contribution 1**: Through efficient self-supervised video pre-training utilizing a masking strategy, we developed a powerful echocardiogram vision encoder that processes entire video clips instead of individual image frames. Notably, all data used for the pre-training stage is sourced from publicly available datasets.
- **Contribution 2**: We introduced a lightweight and innovative spatiotemporal feature fusion network, enhancing predictive accuracy across multiple cardiac clinical tasks.
- **Contribution 3**: Our proposed model surpassed a range of existing state-of-the-art models in echocardiogram analysis, establishing a new benchmark for future research in this field.

2 Related work

2.1 Automated clinical diagnosis for echocardiogram

Significant advancements have been made in automated clinical diagnosis using echocardiograms, largely driven by developments in artificial intelligence (AI) and machine learning (ML). Numerous studies [23, 25, 32, 33] have demonstrated the potential of deep learning models, particularly convolutional neural networks (CNNs), to accurately identify and quantify cardiac structures, assess cardiac function, and detect pathologies in echocardiographic data. For instance, research indicates that AI can automate the measurement of standard echocardiographic parameters, such as left ventricular volume and ejection fraction, achieving accuracy comparable to that of experienced clinicians. Moreover, innovative approaches utilizing recurrent neural networks (RNNs) [34, 35] and 3D CNNs [36] have been explored to better handle the temporal and volumetric data inherent in echocardiogram videos, [24], enhancing the ability to monitor dynamic changes over time. These technologies not only promise to alleviate the workload of cardiologists by automating routine tasks but also aim to standardize echocardiographic assessments, thereby reducing inter-observer variability and improving diagnostic consistency. This growing body of work signifies a pivotal shift toward more sophisticated, AI-driven diagnostic tools in cardiology, suggesting a promising future for the automation of echocardiographic analysis.

The increasing availability of large echocardiogram datasets has further facilitated the development of AI-based diagnostic approaches. [37] Both traditional machine learning models and contemporary deep learning techniques have shown promise in automating echocardiogram interpretation and accurately detecting various heart conditions.[37] However, these models often depend on large amounts of labeled data, which can be expensive and time-consuming to acquire.[3] While supervised deep learning models have demonstrated success, annotating medical videos, such as echocardiograms, requires specialized domain expertise, where even minor errors can lead to significant diagnostic inaccuracies.[38] This bottleneck in generating large, annotated datasets limits the full potential of these models. Consequently, a major challenge within the AI community is to develop methods that can leverage unlabeled data to create robust models without extensive manual labeling.[39]

2.2 Video representation learning

In the deep learning and computer vision communities, the traditional approach to vision analysis and representation learning has predominantly involved fully supervised learning.[40, 27] Training effective models typically necessitates a substantial amount of labeled data, governed by various task-specific labels. When transferring models to other tasks, the prediction head is removed from the pre-trained model, and a new head is established for the new tasks. Recently, semi-supervised video representation learning has gained attention, allowing unlabeled training samples to be supervised using representations from labeled samples.[41] However, the top-down training paradigm employed in supervised or semi-supervised representation learning does not adequately explore the fundamental structure of video data.

Several multi-modal contrastive learning algorithms have been developed to extract video representations from loosely structured text supervision.[42, 43] With the emergence of self-supervised video learning, large-scale foundation models have begun to be built solely using unlabeled data, which is significantly less costly than acquiring high-quality labeled data. Prior knowledge of temporal information is often leveraged to design pretext tasks for self-supervised video learning (SSVP). [44, 45, 46] Contrastive learning [47, 48, 49, 50] Contrastive learning has emerged as a prominent method for enhancing visual representations, although it typically requires substantial data augmentation and large mini-

batch sizes.[51] Researchers have employed CNNs or LSTMs to predict video clips in pixel space, while VideoMAE utilizes a basic masked autoencoder with modern ViT backbones for data-efficient SSVP.

Given the scarcity of labeled echocardiogram datasets, employing traditional fully-supervised models for feature extraction is often impractical, thereby creating a niche for self-supervised models. In the medical domain, certain predetermined data modalities, such as chest X-rays and echocardiograms, exhibit high structural consistency across different patient cases, with significant variation primarily in regions of interest (ROIs) reflecting various health conditions or disease severities. This characteristic of medical visual data stands in contrast to natural visual data, which can vary dramatically even within the same category. Consequently, variational autoencoders (VAEs) [52, 53, 54] are well-suited for modeling echocardiogram data, as they effectively capture intra-class characteristics and compress high-dimensional data into lower-dimensional representations by reconstructing input data from learned latent vectors. A substantial body of research has focused on VAEs for learning representations [55, 56, 57], image generation [58, 59] and this line of inquiry continues to evolve. Inspired by this work, we employed video masked autoencoders (VideoMAE) [26] for learning compact, informative video representations.

2.3 Spatiotemporal feature funsion

Several approaches have been proposed to effectively fuse spatial and temporal features in video understanding tasks. One of the earliest and most influential methods is the Two-Stream Network, which processes spatial features from individual RGB frames using a CNN and temporal motion features from stacked optical flow with another CNN. [60, 61]. This method processes spatial features from individual RGB frames using a CNN and temporal motion features from stacked optical flow using another CNN. These two streams are subsequently fused to leverage both appearance and motion information for tasks such as action recognition. The advent of 3D Convolutional Networks (3D CNNs), exemplified by C3D, marked a significant shift towards integrated spatial-temporal processing.[62, 63, 64] 3D CNNs extend traditional 2D convolutions into the temporal domain, enabling the simultaneous capture of spatial and temporal features from raw video data. More recently, transformer-based models have emerged as powerful tools for spatial-temporal fusion, applying multi-scale self-attention over hierarchical spatial-temporal windows to capture both local and global dependencies across space and time.[65, 66, 67] Our STFF-Net builds upon the Two-Stream Network framework, utilizing both 2D and 3D CNNs along with self-attention blocks to enhance performance in video understanding tasks.

3 Methodology

In this section, we first explain our choice of VideoMAE as the self-supervised model framework and analyze its architecture and properties. We then introduce our pre-trained framework, termed Echo-VideoMAE. Finally, we propose a novel, lightweight SpatioTemporal Feature Fusion Network (STFF-Net) designed to optimize the utilization of representations from the pre-trained video encoder while leveraging spatiotemporal correlations in a low-dimensional representation space.

3.1 Echo-VideoMAE

We utilized the powerful self-supervised video learning framework, Echo-VideoMAE, for our first-stage pre-training.[26] This framework is based on VideoMAE, which was trained on approximately 200,000 echocardiogram samples from the publicly available MIMIC-IV-ECHO dataset.[68] Following the original model's structure, we implemented a straightforward tube masking strategy with a high masking ratio to facilitate MAE pre-training using an asymmetric encoder-decoder architecture derived from ImageMAE.[55] Figure 1(a) illustrates the workflow of the proposed EchoVisionFM framework. Below, we revisit the key components and concepts of VideoMAE.

Masking strategy: As present by VideoMAE, our framework employed high-ratio masking strategy (masked ratio up to 80%-90%) to account for the temporal redundancy in videos.[69, 70] An echocardiogram video typically comprises a series of captured frames that gradually change in the temporal dimension.[70, 71] As shown in Figure 1(b), there is considerable duplication between frames, leading to two significant challenges for masked video auto-encoding. First, maintaining the original temporal frame rate during pre-training would be less effective, necessitating a focus on static or slow motions. Second, the presence of temporal redundancy dilutes motion-based representations, making it relatively straightforward to reconstruct missing pixels using a standard masking ratio (e.g., 50% to 75%). Consequently, dynamic motion representations may not be effectively captured by the encoder backbone.

To address the temporal redundancy observed between consecutive echocardiogram frames, we employed a strided temporal sampling technique for more efficient video pre-training. We randomly selected a video clip with T_{raw} consecutive frames from the original video V, compressing it to T frames using temporal sampling, with each frame

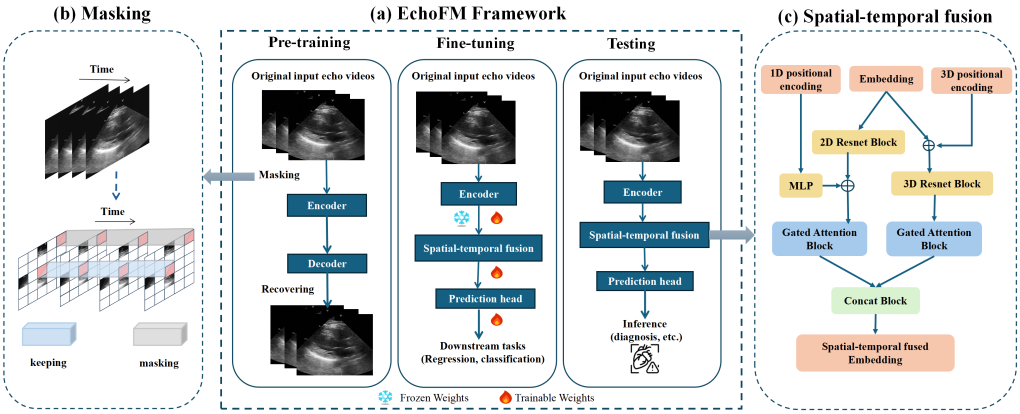


Figure 1: Framework for Echocardiogram Videos Vision Foundation Model (a) The EchoVisionFM framework consists of three stages: pre-training the Echo-VideoMAE, fine-tuning for specific downstream tasks, and testing for inference. During the fine-tuning phase, the encoder weights can either be frozen or further optimized to enhance performance.(b) Tube Masking employs a consistent masking map across all frames, utilizing a high-ratio masking strategy to improve robust feature extraction.(c) The STFF-Net architecture features dual pathways for feature fusion: the joint space-time pathway, which integrates dense absolute 3D positional encoding with a ResNet block, and the disjoint space-time pathway, which processes these dimensions using sparse relative 1D time-series positional encoding.

containing $3 \times H \times W$ pixels. In our experiments, the stride τ was set to 2 for the MIMIC-IV-ECHO dataset. In Echo-VideoMAE, we used a joint space-time cube embedding approach [3, 22, 39], treating each $2 \times 16 \times 16$ cube as a single token embedding. This cube embedding layer generates $\frac{T}{2} \times \frac{H}{16} \times \frac{W}{16}$ 3D tokens, which are then transferred to the channel dimension D. This design reduces spatiotemporal redundancy in frames and enhances the effectiveness of the reconstruction task.

Backbone: Due to the high masking ratio, the encoder receives only a limited number of tokens as input. To capture high-level spatiotemporal information from the unmasked tokens, we employed a vanilla ViT backbone[30] and joint space-time attention [9, 72] for both the encoder and decoder. The multi-head self-attention layer facilitates interactions among all pairs of tokens.[73]

Learning objective: Following the standard workflow of auto-encoders, [52, 58], our Echo-VideoMAE model compresses input data into manifold representations through an encoder and then reconstructs these representations back to the raw pixel space via a decoder. Specifically, VideoMAE performs masking and reconstruction tasks by splitting the input video clip $V \in R^{T \times 3 \times H \times W}$ is splitted into non-overlapping cubes of size $2 \times 16 \times 16$ (stride τ set to be 2). Each cube is treated as a token embedding, with (90%) randomly unmasked tokens supplied to the ViT encoder (Φ_{enc}). The reconstruction is performed using a shallow decoder (Φ_{dec}) applied to the visible tokens and learnable masks. The loss function calculates the mean squared error (MSE) between the normalized masked tokens and the reconstructed ones in pixel space:

$$\mathcal{L}_{recon} = \frac{1}{M} \sum_{i \in M} |V(i) - \widehat{V}(i)|^2 \tag{1}$$

where i is the token index, M is the set of masked tokens, V is the input video clip, and $\widehat{V}$ is the reconstructed version.

By optimizing this loss function, the encoder and decoder are effectively trained to encode input videos and decode manifold representations back to the input data. After pre-training, we retain the lightweight decoder while using the powerful encoder to obtain representations. Unlike traditional CNN-based networks, our ViT encoder is flexible and insensitive to the number of token embeddings, allowing us to input any unmasked video data to generate complete, expressive video representations in a low-dimensional space, suitable for various downstream tasks requiring minimal fine-tuning..

3.2 STFF-Net

We propose a novel, efficient **S**patio**T**emporal **F**eature **F**usion network (**STFF-Net**), designed to leverage all available representations and their spatiotemporal correlations from the pre-trained Echo-VideoMAE..

Inefficient utility of representations: As previously discussed, we utilized a vanilla ViT as the video encoder to transform video clips into a sequence of manifold representations containing the most expressive information from the raw video. However, due to the use of joint space-time attention within the ViT, these learned representations primarily capture global, high-level visual information while neglecting the spatiotemporal relationships between individual video patches.[66]In natural language processing tasks, [73, 74], the relationships between words in a sentence are sequentially one-dimensional, allowing each word to be uniformly attended to by all others. In contrast, the three-dimensional correlations among video patches are crucial and should not be overlooked or simplified through pooling operations, as each patch's relative position significantly impacts the information it conveys.

Architecture: To address these challenges, we designed a customized spatiotemporal feature fusion network, as illustrated in Figure 1(c). This network takes a sequence of patch embeddings from the video encoder as input and outputs a final embedding for downstream tasks, such as regression and classification. To account for the distinct yet complementary relationships between spatial and temporal dimensions, we propose two separate feature fusion pathways for deeper integration of features.

Joint space-time feature fusion pathway: The manifold representations, shaped as $N \times L \times D$ (where N is the mini-batch size, L is the length of patch embeddings, and D is the feature dimension), are first reshaped into a three-dimensional feature map of size $N \times D \times t \times h \times w$, where $t = \frac{T}{\tau}, h = \frac{H}{16}, w = \frac{W}{16}$. Note that $L = t * h * w$. We then add dense 3D positional encoding [75, 76] to the feature maps before passing them into a ResNet block consisting of 3D convolutional layers, batch normalization, and ReLU activation. This strengthens the 3D relationships among the representations. The resulting feature map is flattened while maintaining the batch size and feature dimension. Finally, we apply a gated attention mechanism[77] to assign self-attention weights to each patch embedding, achieving a compact joint space-time video representation, yielding an output size of $N \times D$.

Disjoint space-time feature fusion pathway: In contrast to the previous pathway, we treat the manifold representations as a 2D feature map of size $N * t \times D \times h \times w$. This approach allows us to isolate spatial and temporal relationships to learn disjoint representations. This 2D feature map is processed through a 2D ResNet block for spatial information extraction[29], followed by a large-stride 2D convolutional layer (with a stride of (h, w)) to compress the feature map and capture abstract semantic representations in 2D space. The output is reshaped into a sequential tensor of size $N \times t \times D$. Given that downsampling frames from the original echocardiogram video results in large temporal gaps, these frame indices are encoded using 1D positional encodings generated by a sinusoidal function and transformed by the proposed sparse positional encoding network(a two-layer fully connected network).[73, 78, 79] We then incorporate the learned sparse temporal positional encodings into the sequential tensor, followed by compression of the sequence of embeddings using a gated attention mechanism. The output of this pathway is also of size $N \times D$.

The embeddings from both pathways are concatenated along the feature dimension, resulting in a final embedding of size $N \times 2 * D$. This representation serves as input for any task-specific head. Our experiments demonstrate that the proposed network performs effectively, surpassing state-of-the-art models across multiple downstream tasks.

3.3 Study Approval

This study exclusively used publicly available data.

4 Experiments

In this section, we first introduce the datasets used to pre-train our Echo-VideoMAE and to fine-tune it, with and without the proposed STFF-Net, for comprehensive cardiac assessment. We then present the main results from EchoFM on downstream tasks, comparing its performance to that of vanilla VideoMAE [26], Video ResNet [80], and Vivit [72].

4.1 Datasets

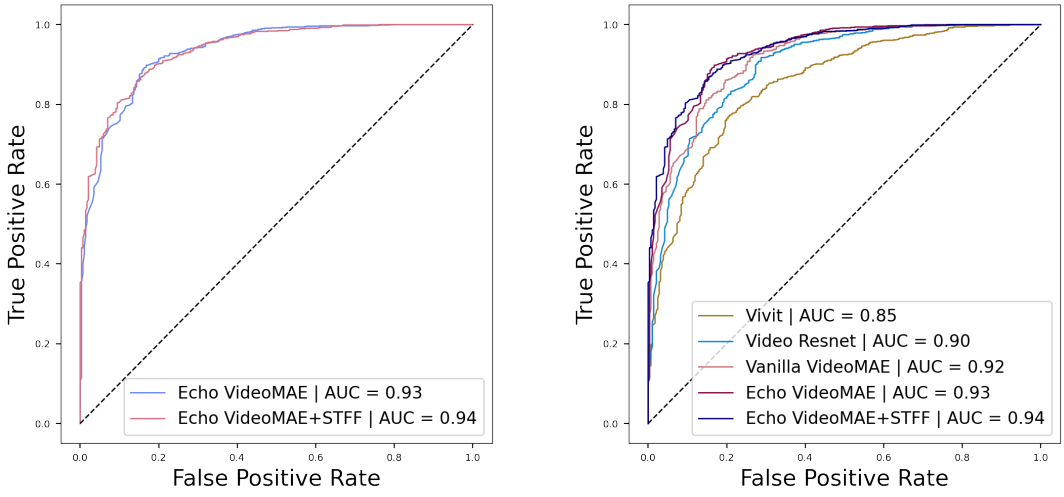
MIMIC-IV-ECHO: We utilized the MIMIC-IV-ECHO dataset [68] for pre-training our Echo-VideoMAE, which is part of the larger MIMIC-IV database. [81] This dataset comprises approximately 500,000 echocardiogram videos collected from 7,243 studies involving 4,579 distinct patients at Beth Israel Deaconess Medical Center between 2017 and 2019. The videos were originally stored as DICOM files, a standard format in medical imaging that includes both video data and metadata. For our purposes, these DICOM files were converted to AVI format using the PyDICOM library, retaining only the video frames and excluding metadata. To preserve comprehensive visual information, we uniformly sampled 16 frames from each original AVI file, resizing each frame to 224×224. Simple data augmentation techniques, such as random vertical and horizontal flips, were applied to enhance data utility and model generalizability. Consequently, all video data used for pre-training had a uniform shape of $16 \times 3 \times 224 \times 224$, with each pixel represented in RGB color

space. To the best of our knowledge, this work is the first to leverage the MIMIC-IV-ECHO dataset—a fully public, de-identified, large database—to establish a video vision foundation model. For practical reasons related to storage, we utilized 40% of the available data from this dataset.

EchoNet Dynamic: The EchoNet-Dynamic database [17] includes 10,030 labeled echocardiogram videos along with human expert annotations, including measurements, tracings, and calculations. Specifically, this dataset comprises apical-4-chamber echocardiogram videos from individuals who underwent imaging as part of routine clinical care at Stanford University Hospital between 2016 and 2018. The original standardized 112x112 pixel videos were upsampled to 224×224 using bilinear interpolation. Each video is linked to clinical measurements and calculations performed by a registered sonographer and verified by an echocardiographer as part of the standard clinical workflow. Key metrics, including left ventricular ejection fraction (LVEF), left ventricular end-systolic volume (ESV), and left ventricular end-diastolic volume (EDV), are associated with each video.

4.2 Echocardiogram interpretation

Using the well-trained echo video encoder from Echo-VideoMAE, we cascaded this encoder with STFF-Net and a classification head, fine-tuning the entire model across a variety of benchmark echocardiogram datasets. We evaluated our method on the held-out test set from EchoNet-Dynamic.[17] The results demonstrated an accuracy of 0.8912, an F1 score of 0.8912, and an area under the ROC curve (AUC) of 0.9323. Figure 2(a) presents the ROC curve for Echo-VideoMAE without STFF, while Figure 2(b) and Table 1 compare Echo-VideoMAE against other baseline models, highlighting the improved performance of Echo-VideoMAE with STFF, which achieved a higher AUC of 0.94.



(a) Echo VideoMAE(w/o STFF) **(b)** Baseline Comparison

Figure 2: Comparison of ROC Curves of EF-Classification

Table 1: Classification performance of LVEF (threshold 50%) on Echonet-Dynamic dataset

	Accuracy	F1 Score	AUC
Vivit	0.8371	0.9013	0.8540
Video ResNet	0.8747	0.9223	0.8975
Vanilla VideoMAE	0.8768	0.9228	0.9136
Echo VideoMAE with max pooling	0.8896	0.9311	0.9330
Echo VideoMAE with STFF-Net	**0.8912**	**0.9323**	**0.9364**

The confusion matrices for the baseline models are displayed in Figure 3, demonstrating that Echo-VideoMAE with STFF-Net outperforms the others in terms of true positives and false positives.

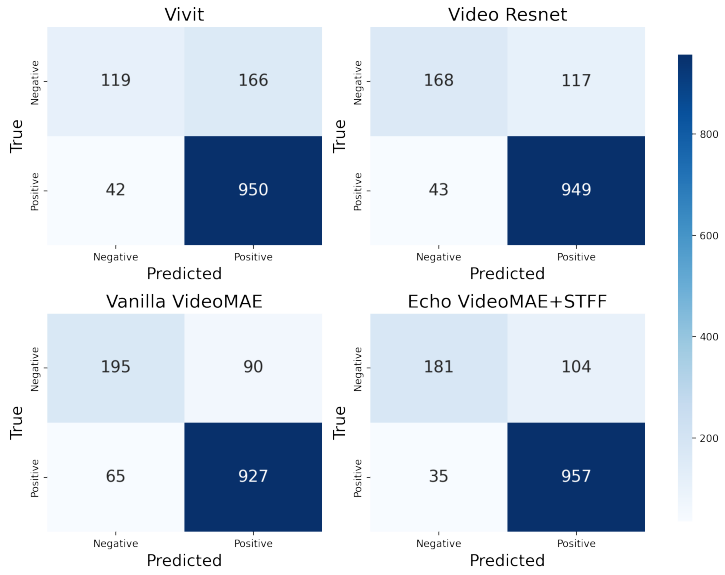


Figure 3: Comparative Confusion Matrix for Baseline Model.

4.3 Cardiac function and pressure assessment

Following the same pipeline, we further evaluated our method on quantitative tasks, specifically estimating left ventricular ejection fraction (LVEF), left ventricular end-systolic volume (ESV), and left ventricular end-diastolic volume (EDV) from EchoNet-Dynamic.[17]

EchoFM predicts LVEF with a mean absolute error (MAE) of 4.18% and a R of 0.8022. The performance metrics are summarized in Tables 2 3 4, indicating that our model exhibits superior performance compared to previous models. Figure 4 illustrates the advantages of the Echo-VideoMAE model, particularly when enhanced by the STFF-Net module, in accurately predicting crucial cardiac functional parameters.

Table 2: Regression performance of LVEF on Echonet-Dynamic dataset

	MAE	MSE	RMSE	R^2
Vivit	0.0668	0.0082	0.0906	0.4510
Video ResNet	0.0544	0.0055	0.0741	0.6323
Vanilla VideoMAE	0.0505	0.0045	0.0671	0.6986
Echo VideoMAE with max pooling	0.0481	0.0040	0.0635	07304
Echo VideoMAE with STFF-Net	**0.0418**	**0.0030**	**0.0547**	**0.8022**

Table 3: Regression performance of ESV on Echonet-Dynamic dataset

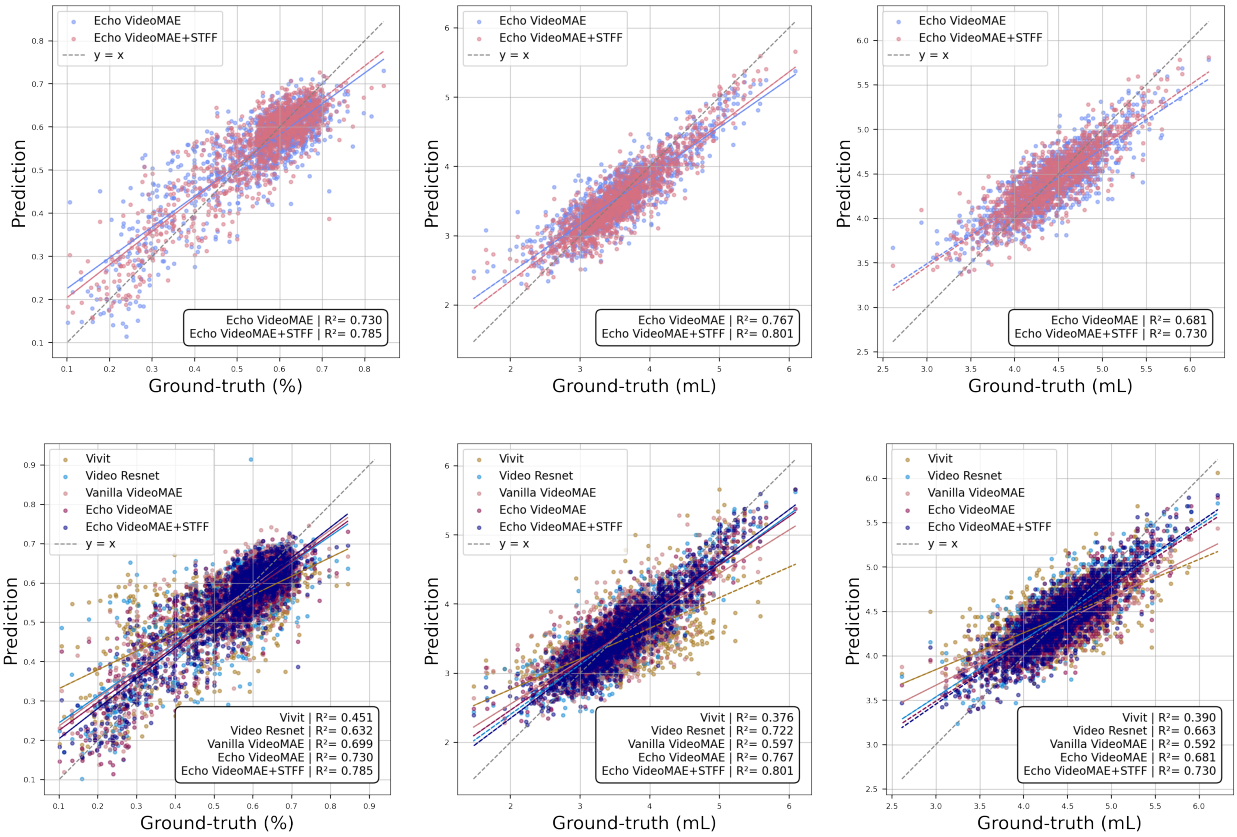
	MAE	MSE	RMSE	R^2
Vivit	0.3704	0.2286	0.4781	0.3763
Video ResNet	0.2473	0.1018	0.3190	0.7223
Vanilla VideoMAE	0.2990	0.1477	0.3843	0.5970
Echo VideoMAE with max pooling	0.2286	0.0856	0.2952	0.7666
Echo VideoMAE with STFF-Net	**0.2096**	**0.0731**	**0.2703**	**0.8006**

5 Discussion

This study demonstrates that large-scale video datasets of echocardiogram studies can serve as a foundation for training and deploying medical video models. Our echocardiogram video foundation model, along with the designed two-stage

Table 4: regression performance of EDV on Echonet-Dynamic dataset

	MAE	MSE	RMSE	R^2
Vivit	0.2732	0.1219	0.3491	0.3901
Video ResNet	0.2009	0.0674	0.2596	0.6626
Vanilla VideoMAE	0.2252	0.0816	0.2856	0.5918
Echo VideoMAE with max pooling	0.1977	0.0638	0.2526	0.6806
Echo VideoMAE with STFF-Net	**0.1796**	**0.0540**	**0.2325**	**0.7296**



(a) LVEF **(b)** ESV **(c)** EDV

Figure 4: Comparison of Echo-VideoMAE with and without the STFF module (top row) alongside baseline models (bottom row) for predicting LVEF, ESV, and EDV. Performance is assessed using the correlation coefficient (R) in relation to the ground truth.

pipeline, effectively addressed several benchmark prediction and assessment tasks. The model exhibited remarkable transferability and robustness in domain adaptation, having been pre-trained on data from one healthcare facility and fine-tuned on downstream tasks from entirely separate hospitals with diverse acquisition settings. Moreover, the Echo-video encoder, leveraging self-supervised video pre-training and a unique masking strategy, adapted well to the structured manifold space, yielding strong and transferable visual representations. Through the use of SSVP and masked auto-encoding, our model outperformed existing state-of-the-art baselines, which often relied on fully supervised learning. Additionally, the proposed STFF-Net, designed to leverage the spatiotemporal correlations within video representations, significantly enhanced performance across nearly all assessments of cardiac function and key clinical value predictions.

EchoVisionFM is the first echocardiogram video foundation model trained exclusively on publicly available datasets. In contrast, previous echo-related vision models were trained on private datasets from specific healthcare systems,

which hampers the broader application of these AI systems. [82, 23, 83] Notably, EchoVisionFM outperformed previous models, achieving competitive results. For instance, it attained a mean absolute error (MAE) of 4.18% in external validation of LVEF prediction from Echonet-Dynamic, while the most recent video-based LVEF AI model [83] achieved an MAE of 4.34% and a multi-modal AI model [82] reported an MAE of 7.1%. For the clinically significant LVEF threshold of 50%, EchoVisionFM surpassed previous AI models, achieving an AUC of 0.9364 compared to the 0.89–0.90 achieved by others. [82].

A central challenge in integrating emerging AI systems in the medical field is the scarcity of available training data. Previous echocardiogram AI models were typically trained on a maximum of 150,000 echocardiogram videos.[84] Medical labeling remains a labor-intensive task, even for experienced clinicians, often requiring hundreds or thousands of labeled samples for challenges like regression and segmentation.[85, 86, 87] Consequently, training a medical foundation model using fully supervised learning is often impractical, despite its potential for superior performance due to robust supervision. The first-stage pre-training utilized in this study capitalized on large unlabeled public datasets, instilling comprehensive video prior knowledge into our Echo-video encoder. The well-trained encoder can serve not only as a feature extractor for echocardiogram vision tasks related to heart function evaluation but also as a visual encoder for future language-vision models. We believe that this out-of-the-box medical video encoder has significant potential to enable more powerful multi-modal foundation models capable of processing various 3D medical video data, including CT scans, MRIs, and endoscopy videos.

In cardiovascular diagnosis and evaluation, clinicians typically use echocardiogram videos to assess patient conditions and severity. Over the past few years, numerous vision AI models have been developed, exploring both pure vision models and language-vision models.[5, 6, 7, 8, 9, 10, 3, 13] and language-vision models [82, 88] We noted that many echocardiogram-related models rely on 2D image frames extracted from the entire echocardiogram[6, 82, 23] rather than directly modeling 3D video clips, which contain crucial motion-based information for accurately analyzing heart contraction and function. Our model processes entire echocardiogram videos as input, ensuring that comprehensive visual information is utilized while avoiding significant computational costs. Intuitively, video-based features are likely to be more informative than image-based features, and empirically, our model has indeed outperformed previous AI models on comparable visual tasks.

In addition to the Echo-video encoder, we developed a novel, simplified Spatiotemporal Feature Fusion Network (STFF-Net). This network is integrated with the Echo-video encoder and fine-tuned during the second stage. For standard ViT encoders using the original self-attention mechanism, [73, 30], adding a learnable classification token [30] pooling the last-layer token embeddings is common practice for specific downstream tasks. However, this approach overlooks the correlations between video patches and fails to leverage the full range of information available from the echocardiogram videos. Our design can be adapted to any ViT-style encoder, potentially enhancing overall performance.

Limitation

Despite the promising results demonstrated by our Echo-VideoMAE model and STFF-Net, there are several limitations that warrant discussion First, while our model was trained on the MIMIC-IV-ECHO dataset and fine-tuned on the EchoNet-Dynamic database, the generalizability of our approach to diverse populations and varying clinical practices remains uncertain. The datasets used primarily originate from specific healthcare facilities, which may not capture the full spectrum of echocardiographic variations encountered in broader clinical settings. Second, although we utilized a high masking ratio to exploit temporal redundancy in echocardiogram videos, our approach may overlook finer temporal details crucial for specific cardiac assessments. The sampling and compression of video frames may result in the loss of critical motion-related information, potentially affecting the accuracy of certain predictions. Third, given the relatively limited diversity of echocardiographic data in the datasets used, there is a risk of overfitting. While the pre-training strategy mitigates this concern, continuous monitoring and validation on external datasets are necessary to ensure robust performance across different clinical scenarios..

Future work

Future directions for this work involve enhancing the spatiotemporal network to capture more intricate patterns in heart dynamics, integrating additional modalities such as electrocardiograms, and deploying the system in real-time clinical settings. Additionally, exploring multi-task learning to predict multiple cardiac conditions simultaneously could significantly enhance its utility for comprehensive heart disease detection. Like many deep learning models, Echo-VideoMAE operates as a "black box," which can complicate the interpretation of its decision-making processes. Gaining a clearer understanding of how the model generates its predictions is essential for fostering clinical adoption and building trust. Therefore, further research into interpretability techniques will be critical. As AI continues to

advance, EchoVisionFM offers a promising foundation for developing more efficient, scalable, and accurate diagnostic tools in medicine.

6 Conclusion

In this study, we developed EchoVisionFM, a pioneering echocardiogram video foundation model that leverages large-scale, publicly available datasets for training and deployment. Our approach demonstrates significant advancements in the prediction and assessment of cardiac function through a self-supervised learning framework, highlighting the model's robust transferability and adaptability across diverse clinical settings. By integrating Echo-VideoMAE with a novel STFF-Net, we achieved superior performance in various cardiac assessment tasks compared to existing state-of-the-art models. The findings suggest that utilizing a complete echocardiogram video rather than isolated frames provides richer information, leading to more accurate predictions of critical cardiac parameters. EchoVisionFM represents a significant step forward in harnessing the potential of AI in cardiology, paving the way for more effective diagnostic tools that can enhance patient care and outcomes. As we continue to innovate in this space, we aim to contribute to the broader goal of advancing healthcare through intelligent, data-driven solutions.

Acknowledgements The authors would like to thank Northwestern's High-Performance Computing (HPC) cluster, and the data storage and computational resources support from Quest.

Funding This study is partially supported by the American Heart Association Grant (24GWTGTG1268589).

Contribution Statement JY and ZZ designed the study. ZZ and JY contributed to data analyses. ZZ, SD, QW and JY contributed to the writing of the manuscript. XW and QW contributed to data management. All authors read and approved the final version of the manuscript.

Data Availability Statement MIMIC-IV-ECHO data are available at Medical Information Mart for Intensive Care: `https://physionet.org/content/mimic-iv-echo/0.1/`. Echonet-Dynamic dataset is available at `https://aimi.stanford.edu/datasets/echonet-dynamic-cardiac-ultrasound`. The relevant code and analyses are available at: `https://github.com/ZiyangZhang0511/echo_vision`.

Conflict of Interest Statement None.

References

[1] 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.

[2] NAGUEH SF. Recommendations for the evaluation of left ventricular diastolic function by echocardiography. *J Am Soc Echocardiogr*, 22:107–133, 2009.

[3] Jeffrey Zhang, Sravani Gajjala, Pulkit Agrawal, Geoffrey H Tison, Laura A Hallock, Lauren Beussink-Nelson, Mats H Lassen, Eugene Fan, Mandar A Aras, ChaRandle Jordan, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation*, 138(16):1623–1635, 2018.

[4] Yeonyee E Yoon, Sekeun Kim, and Hyuk-Jae Chang. Artificial intelligence and echocardiography. *Journal of Cardiovascular Imaging*, 29(3):193, 2021.

[5] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

[6] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

[7] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[8] 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.

[9] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[10] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[11] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7:25, 2020.

[12] Ali Madani, Jia Rui Ong, Anshul Tibrewal, and Mohammad RK Mofrad. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. *NPJ digital medicine*, 1(1):1–11, 2018.

[13] Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H Chen, Robert A Harrington, David H Liang, Euan A Ashley, and James Y Zou. Deep learning interpretation of echocardiograms. *NPJ digital medicine*, 3(1):10, 2020.

[14] Tawsifur Rahman, Muhammad EH Chowdhury, Amith Khandakar, Khandaker R Islam, Khandaker F Islam, Zaid B Mahbub, Muhammad A Kadir, and Saad Kashem. Transfer learning with deep convolutional neural network (cnn) for pneumonia detection using chest x-ray. *Applied Sciences*, 10(9):3233, 2020.

[15] Lhuqita Fazry, Asep Haryono, Nuzulul Khairu Nissa, Naufal Muhammad Hirzi, Muhammad Febrian Rachmadi, Wisnu Jatmiko, et al. Hierarchical vision transformers for cardiac ejection fraction estimation. In *2022 7th International Workshop on Big Data and Information Security (IWBIS)*, pages 39–44. IEEE, 2022.

[16] Bohan Liu, Hao Chang, Dong Yang, Feifei Yang, Qiushuang Wang, Yujiao Deng, Lijun Li, Wenqing Lv, Bo Zhang, Liheng Yu, et al. A deep learning framework assisted echocardiography with diagnosis, lesion localization, phenogrouping heterogeneous disease, and anomaly detection. *Scientific Reports*, 13(1):3, 2023.

[17] David Ouyang, Bryan He, Amirata Ghorbani, Matt P Lungren, Euan A Ashley, David H Liang, and James Y Zou. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In *NeurIPS ML4H Workshop*, pages 1–11, 2019.

[18] Xin Liu, Yiting Fan, Shuang Li, Meixiang Chen, Ming Li, William Kongto Hau, Heye Zhang, Lin Xu, and Alex Pui-Wai Lee. Deep learning-based automated left ventricular ejection fraction assessment using 2-d echocardiography. *American Journal of Physiology-Heart and Circulatory Physiology*, 321(2):H390–H399, 2021.

[19] Mahboobeh Jafari, Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Parisa Moridian, Roohallah Alizadehsani, Abbas Khosravi, Sai Ho Ling, Niloufar Delfan, Yu-Dong Zhang, et al. Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review. *Computers in Biology and Medicine*, 160:106998, 2023.

[20] Fumin Guo, Matthew Ng, Idan Roifman, and Graham Wright. Cardiac magnetic resonance left ventricle segmentation and function evaluation using a trained deep-learning model. *Applied Sciences*, 12(5):2627, 2022.

[21] Andreas Leha, Kristian Hellenkamp, Bernhard Unsöld, Sitali Mushemi-Blake, Ajay M Shah, Gerd Hasenfuß, and Tim Seidler. A machine learning approach for the prediction of pulmonary hypertension. *PloS one*, 14(10):e0224453, 2019.

[22] Maryam Alsharqi, WJ Woodward, JA Mumith, DC Markham, Ross Upton, and Paul Leeson. Artificial intelligence and echocardiography. *Echo Research & Practice*, 5(4):R115–R125, 2018.

[23] 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.

[24] James P Howard, Jeremy Tan, Matthew J Shun-Shin, Dina Mahdi, Alexandra N Nowbar, Ahran D Arnold, Yousif Ahmad, Peter McCartney, Massoud Zolgharni, Nick WF Linton, et al. Improving ultrasound video classification: an evaluation of novel deep learning methods in echocardiography. *Journal of medical artificial intelligence*, 3, 2020.

[25] Shahram Ebadollahi, Shih-Fu Chang, Henry D Wu, and Shin Takoma. Echocardiogram video summarization. In *Medical Imaging 2001: Ultrasonic Imaging and Signal Processing*, volume 4325, pages 492–501. SPIE, 2001.

[26] 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.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[30] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[31] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[32] Rashmi Nedadur, Bo Wang, and Wendy Tsang. Artificial intelligence for the echocardiographic assessment of valvular heart disease. *Heart*, 108(20):1592–1599, 2022.

[33] Kenya Kusunose, Akihiro Haga, Takashi Abe, and Masataka Sata. Utilization of artificial intelligence in echocardiography. *Circulation Journal*, 83(8):1623–1629, 2019.

[34] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[35] I Sutskever. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014.

[36] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[37] Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout. Fast and accurate view classification of echocardiograms using deep learning. *NPJ digital medicine*, 1(1):6, 2018.

[38] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.

[39] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[40] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[41] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10399, 2021.

[42] Tianhao Li and Limin Wang. Learning spatiotemporal features via video and text pair discrimination. *arXiv preprint arXiv:2001.05691*, 2020.

[43] Jonathan C Stroud, Zhichao Lu, Chen Sun, Jia Deng, Rahul Sukthankar, Cordelia Schmid, and David A Ross. Learning video representations from textual web supervision. *arXiv preprint arXiv:2007.14937*, 2020.

[44] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

[45] 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.

[46] 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.

[47] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *European conference on computer vision*, pages 312–329. Springer, 2020.

[48] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in neural information processing systems*, 33:5679–5690, 2020.

[49] 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.

[50] Ali Ghelmani and Amin Hammad. Self-supervised contrastive video representation learning for construction equipment activity recognition on limited dataset. *Automation in Construction*, 154:105001, 2023.

[51] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023.

[52] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[53] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[54] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

[55] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

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

[57] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2506–2517, 2024.

[58] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[59] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[60] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[61] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann. Hidden two-stream convolutional networks for action recognition. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 363–378. Springer, 2019.

[62] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[63] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[64] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[65] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023.

[66] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[67] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3333–3343, 2022.

[68] Brian Gow, Tom Pollard, Nathaniel Greenbaum, Benjamin Moody, Alistair Johnson, Elizabeth Herbst, Jonathan W Waks, Parastou Eslami, Ashish Chaudhari, Tanner Carbonati, et al. Mimic-iv-echo: Echocardiogram matched subset. *PhysioNet https://doi. org/10.13026/EF48-V217*.

[69] Xiangxiang Dai, Peng Yang, Xinyu Zhang, Zhewei Dai, and Li Yu. Respire: Reducing spatial–temporal redundancy for efficient edge-based industrial video analytics. *IEEE Transactions on Industrial Informatics*, 18(12):9324–9334, 2022.

[70] Zhang Zhang and Dacheng Tao. Slow feature analysis for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):436–450, 2012.

[71] Lin Sun, Kui Jia, Tsung-Han Chan, Yuqiang Fang, Gang Wang, and Shuicheng Yan. Dl-sfa: Deeply-learned slow feature analysis for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2632, 2014.

[72] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[73] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[74] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2, 2019.

[75] Changyong Shu, Jiajun Deng, Fisher Yu, and Yifan Liu. 3dppe: 3d point positional encoding for transformer-based multi-camera 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3580–3589, 2023.

[76] Jianqiao Zheng, Sameera Ramasinghe, and Simon Lucey. Rethinking positional encoding. *arXiv preprint arXiv:2107.02561*, 2021.

[77] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[78] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[79] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10041, 2021.

[80] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[81] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[82] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, pages 1–8, 2024.

[83] Adil Dahlan, Cyril Zakka, Abhinav Kumar, Laura Tang, Rohan Shad, Robyn Fong, and William Hiesinger. Echocardiogram foundation model–application 1: Estimating ejection fraction. *arXiv preprint arXiv:2311.12582*, 2023.

[84] Bryan He, Alan C Kwan, Jae Hyung Cho, Neal Yuan, Charles Pollick, Takahiro Shiota, Joseph Ebinger, Natalie A Bello, Janet Wei, Kiranbir Josan, et al. Blinded, randomized trial of sonographer versus ai cardiac function assessment. *Nature*, 616(7957):520–524, 2023.

[85] Jasper Tromp, Paul J Seekings, Chung-Lieh Hung, Mathias Bøtcher Iversen, Matthew James Frost, Wouter Ouwerkerk, Zhubo Jiang, Frank Eisenhaber, Rick SM Goh, Heng Zhao, et al. Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. *The Lancet Digital Health*, 4(1):e46–e54, 2022.

[86] 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.

[87] Emily S Lau, Paolo Di Achille, Kavya Kopparapu, Carl T Andrews, Pulkit Singh, Christopher Reeder, Mostafa Al-Alusi, Shaan Khurshid, Julian S Haimovich, Patrick T Ellinor, et al. Deep learning–enabled assessment of left heart structure and function predicts cardiovascular outcomes. *Journal of the American College of Cardiology*, 82(20):1936–1948, 2023.

[88] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Multimodal foundation models for echocardiogram interpretation. *arXiv preprint arXiv:2308.15670*, 2023.

[89] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[90] Jirka Borovec, William Falcon, Akihiro Nitta, Ananya Harsh Jha, otaj, Annika Brundyn, Donal Byrne, Nathan Raw, Shion Matsumoto, Teddy Koker, Brian Ko, Aditya Oke, Sidhant Sundrani, Baruch, Christoph Clement, Clément POIRET, Rohit Gupta, Haswanth Aekula, Adrian Wälchli, Atharva Phatak, Ido Kessler, Jason Wang, JongMok Lee, Shivam Mehta, Zhengyu Yang, Garry O'Donnell, and zlapp. Lightning-ai/lightning-bolts: Minor patch release, December 2022.

[91] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019.

[92] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. `https://github.com/huggingface/accelerate`, 2022.

[93] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[94] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[95] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.

[96] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Appendix

7 Model architecture

Here, we provide an overview of the model backbone, prediction head, and loss function for all our experiments in Table 5.

Table 5: Model architecture

	backbone	prediction head	loss function
Echo-VideoMAE with STFF	ViT-B	2-layer MLP	mean square erro for regression, binary cross entropy for classification
Echo-VideoMAE	ViT-B		
Vanilla VideoMAE	ViT-B		
Video Resnet	Resnet50		
Vivit	ViT-B		

8 Training setting

We conducted our experiments using four GPUs for pre-training on the MIMIC-IV-ECHO dataset and one GPU for fine-tuning on downstream labeled datasets. To enhance training and inference speed, we utilized the PyTorch [89], PyTorch Lightning [90], Tramsformers [91] and Accelerate [92] frameworks. Additionally, we have made our code and pre-trained models publicly available to support future research in echocardiogram and general medicine.

MIMIC-IV-ECHO. Our Echo-VideoMAE was pre-trained for 50 epochs on the MIMIC-IV-ECHO dataset by default. We employed simple random flip data augmentation with a 0.5 probability. The detailed pre-training configuration is presented in Table 6.

Table 6: Pre-training setting

	MIMIC-IV-ECHO[68]
optimizer	AdamW [93]
learning scheduler	LinearWarmupCosineAnnealing [94]
initial learning rate	1e-4
warmup learning rate	1e-5
optimizer momentum	0.9, 0.95 [95]
weight decay	0.05
warmup epochs	2k
training epochs	50k
batch size	64
filp augmentation	yes

Echonet-Dynamic. Our fine-tuning experiments were conducted over 50 epochs on the downstream dataset. We employed simple random flip data augmentation with a 0.5 probability. The detailed fine-tuning configuration is provided in Table 7. We trained Vivit and Video ResNet from scratch using fully supervised learning. Additionally, we fine-tuned Vanilla VideoMAE with pre-trained weights [26] on Kinetics-400 [96]. The echo video encoder was fine-tuned both with and without the STFF-Net, utilizing the pre-trained weights from our first-stage pre-training.

Table 7: Fine-tuning setting

	Echonet-Dynamic [17]
optimizer	AdamW [93]
learning scheduler	LinearWarmupCosineAnnealing [94]
initial learning rate	5e-4
warmup learning rate	1e-5
optimizer momentum	0.9, 0.999
weight decay	0.01
warmup steps	5k
training epochs	20k
batch size	32
filp augmentation	yes