ECHO-VISION-FM: A PRE-TRAINING AND FINE-TUNING FRAMEWORK FOR ECHOCARDIOGRAM VIDEO 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 essential insights into cardiac health, yet their complex, multidimensional data poses significant challenges for analysis and interpretation. Existing deep learning models for echocardiogram analysis often rely heavily on supervised training, which limits their generalizability and robustness across different datasets and clinical environments.

Objective: To develop and evaluate **Echo-Vision-FM** (**Echo**cardiogram 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. Echo-Vision-FM aims to produce robust and transferable video representations, improving downstream performance across diverse echocardiogram datasets and clinical conditions.

Methods: The proposed framework employs advanced self-supervised video learning through a masked auto-encoding technique, which compresses segments of video data and reconstructs the full video by masking non-overlapping video patches. An asymmetric encoder-decoder architecture underpins this approach. To further enhance the learned representations, we introduce **STF-Net**, a **S**patial-**T**emporal **F**usion **Net**, which integrates spatial and temporal correlations from the video representations. We pre-trained Echo-Vision-FM using the MIMIC-IV-ECHO dataset and fine-tuned it across multiple downstream datasets for specific clinical tasks, including morphological value estimation and the diagnosis of heart function and diseases.

Results: Echo-Vision-FM achieved superior performance in classifying left ventricular ejection fraction (LV_{EF}), with an accuracy of 0.905, an F1 score of 0.941, and an AUC of 0.931. In regression tasks, Echo-Vision-FM outperformed state-of-the-art models, achieving a mean absolute error (MAE) of 3.87% and an r^2 of 0.825 for LV_{EF} prediction. The model also demonstrated significant improvements in estimating end-systolic and end-diastolic volumes, with r^2 values of 0.782 and 0.742, respectively. Incorporating STF-Net further enhanced performance across all tasks.

Conclusion: Our results demonstrate that large-scale self-supervised video learning on echocardiogram data enables the extraction of transferable and clinically relevant features, surpassing existing methods. The Echo-Vision-FM framework, particularly with the inclusion of STF-Net, significantly improves the extraction of spatiotemporal features, resulting in enhanced predictive accuracy for a range of cardiac parameters. Echo-Vision-FM offers a scalable and effective solution for echocardiogram analysis, with promising applications in clinical diagnostics and research.

Keywords: Echocardiogram, Self-supervised video learning, Masked auto-encoding, Spatial-temporal feature fusion, 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 information necessary for assessing cardiac dynamics [23, 24, 25]. Static frames provide only snapshots, omitting continuous motion and temporal 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 and expert models, 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. Although dissecting video data directly is a more powerful way to build VFMs than that of processing single frames, a large number of labeled video data still lacks, which hinder the training and construction of video-VFMs.

To address these aforementioned limitations, we propose two-stage artificial intelligence (AI) model, **Echo-Vision-FM** (**Echo**cardiogram video **Vision F**oundation **M**odel) for automating echocardiogram diagnoses. First we adopt a state-of-the-art self-supervised video learning (SSVL) method, masked auto-encoding [26], that utilize large-scale, unlabeled echocardiogram video data for pre-training a video encoder in echocardiogram domain. This practice relieve the extensive demands to labeled medical data to large extents. After pre-training, we transfer the well-trained **echo-video encoder** to a series of downstream tasks related to heart functions. In addition, to further enhance the learned video representations from the echo-video encoder and boost performance on clinical tasks, we devised a concise yet efficient fusion network, called the **S**patial-**T**emporal **F**usion **Net**work (**STF-Net**). Extensive experiments demonstrated that our model achieves state-of-the-art results in echocardiogram tasks, clearly illustrating the superiority of Echo-Vision-FM over previous methods. In summary, our contributions are as follows:

- We introduce **Echo-Vision-FM**, including a echo-video encoder and a novel representation fusion network. The whole model processes entire video clips instead of individual image frames, which gives deeper insights into human cardiac function.
- We adapt an advanced self-supervised video learning strategy for pre-training a generalizable video encoder, leveraging large-scale unlabeled echocardiograms. To our knowledge this work is the first one that use totally public medical database to build up VFM in echocardiogram.

- We devise a novel and simplified **S**patial-**T**emporal **F**usion **Net**work (**STF-Net**) to further encode the learned video representations for downstream tasks. This design can be easily adapted to any transformer-based video encoder.
- We analyze the impact of different pre-training setups on the pre-trained echo-video encoder and examine the model's capabilities across various downstream tasks. This analysis highlights Echo-Vision-FM's utility in different clinical contexts and offers valuable insights for real-world medical applications.
- We demonstrate that our proposed method outperformed a range of existing state-of-the-art models in echocardiogram analysis, setting 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 AI and machine learning (ML). Numerous studies [23, 25, 27, 28] 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 have indicated 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) [29, 30] and 3D CNNs [31] 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 [32]. Both traditional machine learning models and contemporary deep learning techniques have shown promise in automating echocardiogram interpretation and accurately detecting various heart conditions [32]. 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 [33]. 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 [34]. In contrast, we employed a SSVL method to circumvent the needs of annotated data, shortening the time of building effective VFMs and advancing the frontier of medical AI.

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 [35, 36]. Training effective models typically necessitates a substantial amount of labeled data, governed by various task-specific class labels. When transferring models to other tasks, the simple classifier 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 [37]. However, the top-down training paradigm employed in supervised or semi-supervised representation learning and the shortage of well-annotated video data 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 [38, 39]. 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) [40, 41, 42]. Contrastive learning [43, 44, 45, 46] has emerged as a prominent method for enhancing visual representations, although it typically requires substantial data augmentation and large mini-batch sizes.[47] Researchers have employed CNNs or LSTMs to predict video clips in pixel space, while VideoMAE [26] 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 representation learning 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) [48, 49, 50] are well-suited for modelling echocardiogram data, as they effectively capture intra-class characteristics and compress high-dimensional, complicated data into lower-dimensional representations by reconstructing input data from learned representaions. A substantial body of research has focused on VAEs for learning representations [51, 52, 53], vision generation [54, 55] and this line of works continues to evolve. Inspired by [26], we adapted video masked auto-encodeing for learning compact, informative video representations.

2.3 Spatiotemporal feature fusion

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 [56, 57]. 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 [58, 59, 60]. 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 [61, 62, 63]. Our STF-Net built upon the Two-Stream Network framework, utilizing convolutional operations [64] and gated self-attention mechanism [65] along with spatial positional encodings [66] to compress a sequence of learned video representations into a more compact yet expressive vector.

3 Methodology

In this section, we provided a detailed overview of our two-stage framework, including all its components(see Figure 1). First, in §3.1, we introduced the adapted masked auto-encoding[26] strategy tailored for echocardiogram data, which enables the development of a generalizable echo-video encoder. Next in§3.2, we presented our novel approach for the fusion of steep video representations (see Figure 2). Finally, in §3.3, we discussed the concrete implementations of all the components within the framework.

3.1 Pre-training framework

Due to the limited availability of annotated echocardiographic data, it is not feasible to train a video encoder in a supervised manner that is sufficiently transferable to generalize across related but distinct data domains. Therefore, we turned to self-supervised video learning (SSVL) for effective pre-training. Specifically, we adopted the masked auto-encoding approach [51, 26] to train a video encoder capable of generating compact video representations, as illustrated in 1 **a**. We chose VideoMAE as the foundation for our approach. Below, we discussed two key components of the first-stage pre-training process.

Masking strategy and input embeddings. One of the most critical aspects of masked auto-encoding is its masking strategy, which directly influences the capacity of the learned encoder. The pretext task in this technique involves recovering the original input data from a fraction of it, referred to as the reconstruction task. Therefore, how we select visible parts of the data and mask the remaining portions, making them inaccessible to the model's information flow, is crucial. This approach is particularly well-suited for video processing, as video data often contains significant temporal redundancy [67, 68], whether in natural or medical contexts. In the case of echocardiogram videos, the content typically consists of a series of frames that change gradually over time [68, 69]. This temporal redundancy presents two key 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 movements. Second, the redundancy reduces the complexity of motion-based representations, making it easier to reconstruct missing pixels, which can prevent the encoder from effectively capturing dynamic motion representations.To address this, we followed a standard cube embedding procedure [70, 71] for transforming raw video data into input embeddings, but we reduced the temporal grid stride to be smaller than the 2D spatial grid stride. This adjustment helped to retain richer temporal information for the model.

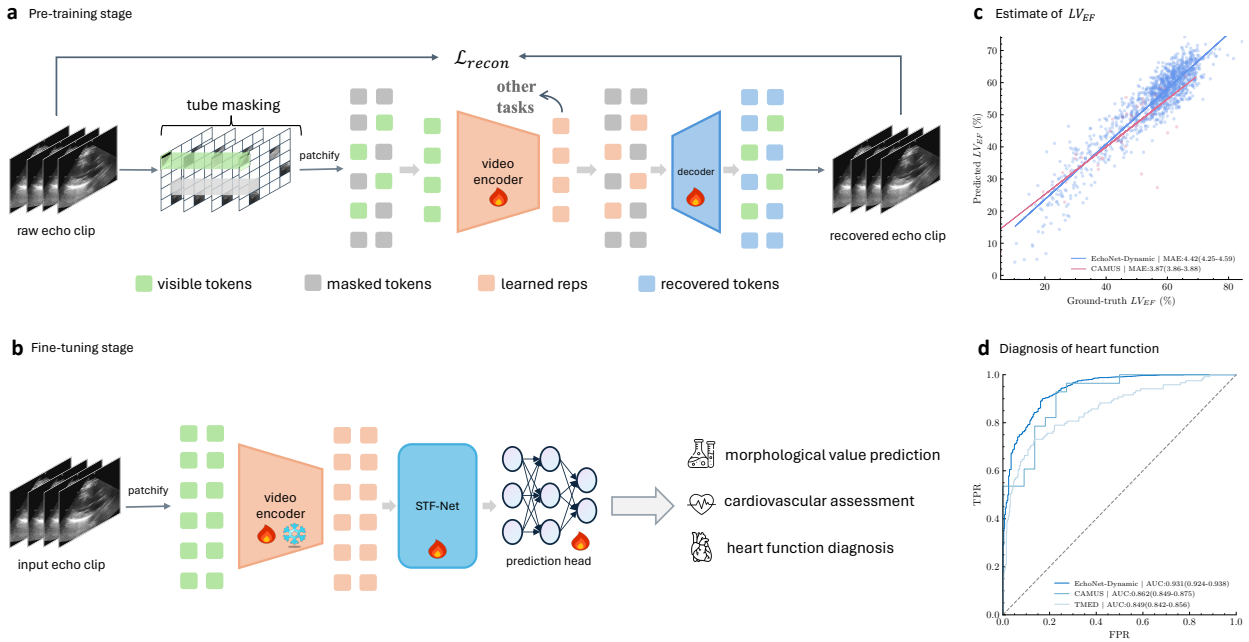

Figure 1: Overview of **Echo-Vision-FM** (Echocardiogram Videos Vision Foundation Model). **a**, A video encoder and decoder were pre-trained using a self-supervised video learning (SSVL) approach, utilizing only fully unannotated video data. We employed an advanced masked auto-encoding technique for SSVL, where the raw video input was reconstructed from a partially visible input using an asymmetric structure and a reconstruction loss. The learned representations generated by the video encoder were then used for downstream tasks. **b**, We discarded the decoder and fine-tuned only the video encoder on specific clinical tasks. To enhance the fine-tuned model's performance, we introduced a novel Spatial-Temporal Fusion Network (STF-Net), which is inserted between the video encoder and any task-specific head. **c**, Scatter plot of predictions versus labels for left ventricular ejection fraction (LV_{EF}) on a held-out test dataset from Stanford Healthcare (Echonet-dynamic; blue, n=10,030) and University Hospital of St. Etienne (CAMUS; red, n=500). **d**, AUC performance for various heart function diagnoses, including LV_{EF} assessment on held-out test datasets from Stanford Healthcare, University Hospital of St. Etienne, and Aortic Stenosis from Tufts Medical Center. FPR: false positive rate; TPR: true positive rate.

In practice, we selected a video clip V with T_{raw} consecutive static frames and compressed it into T frames using temporal uniform sampling. Each frame contains $3 \times H \times W$ pixels in RGB color space. We denoted temporal grid stride as τ and the spatial grid stride as h, w. As a result, the cube embedding layer generated a sequence of token embeddings with length $L_{seq} = \frac{T}{\tau} * \frac{H}{h} * \frac{W}{w}$, which are then mapped to the channel dimension d_{model}. Next, we masked a fraction of tokens based on a designated mask ratio (e.g. 70%), feeding only the unmasked tokens into the encoder backbone. These design choices helped reduce spatiotemporal redundancy in the raw input data and enhanced the effectiveness of the reconstruction task.

Learning objective. Following the standard workflow of auto-encoding [48, 54] and vision transformer-based architectures [72], our pre-training framework embeded input data into a sequence of video representations through an video encoder (Φ_{enc}). These learned representations were then used to recover the downsampled input video via a shallow decoder (Φ_{dec}). The primary learning objective was the common mean squared error (MSE) between the masked tokens and their reconstructed counterparts:

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

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, both the encoder and decoder were trained to effectively encode input videos and decode the learned representations, along with the masked tokens, back to the original data. After pre-training, we discarded the lightweight decoder and used the powerful encoder to extract video representations. Unlike traditional

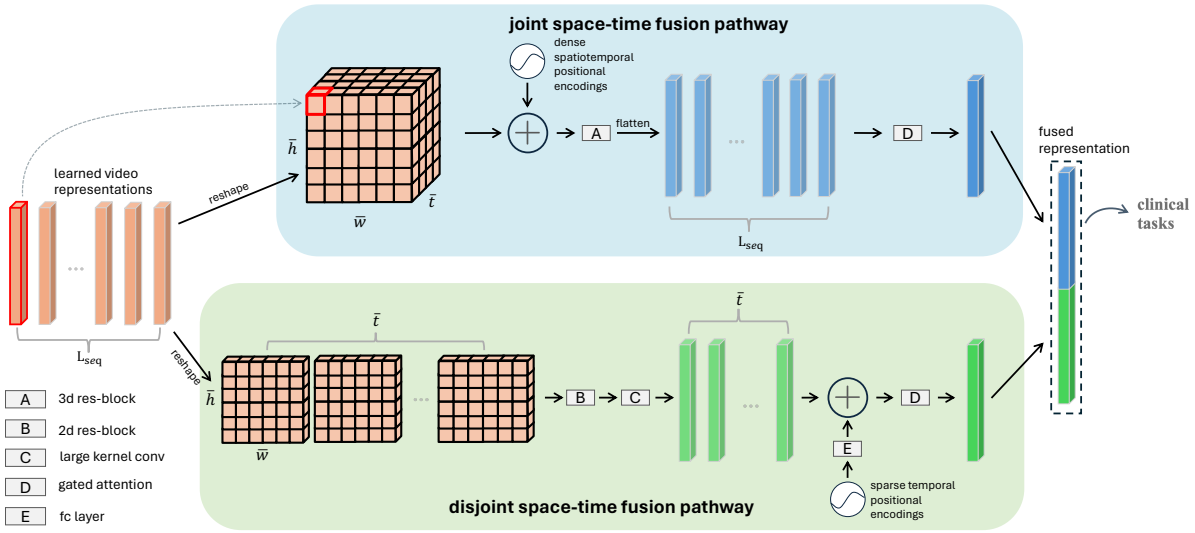

Figure 2: STF-Net takes as input the learned video representations from the pre-trained echo-video encoder and outputs an aggregated fused representation, which is formed by concatenating two feature vectors from dual pathways. For simplicity, the batch dimension is omitted in this illustration. **Top**: The video representations (orange) are reshaped into three-dimensional feature maps, followed by the addition of 3D positional encodings and a ResNet block to generate light-blue vectors. A gated self-attention mechanism then aggregates all the features into a single representation (blue). **Bottom** (disjoint space-time fusion): The video representations are treated as a sequence of 2D feature maps, processed through a 2D ResNet block, and the spatial dimensions are reduced using a large-kernel convolution to form light-green vectors. Then, large-interval, sparse learnable positional encodings are added, and all features are aggregated into a single representation (green). Finally, the blue and green representations are concatenated into one fused representation, which is used for downstream tasks.

CNN-based networks, our vision transformer-based video encoder is highly flexible and not sensitive to the number of token embeddings. This allows us to input any unmasked video data and generate comprehensive, expressive video representations that are suitable for a variety of downstream tasks with minimal fine-tuning.

3.2 Spatial-temporal fusion network

We propose a novel yet concise **S**patial-**T**emporal **F**usion **Net**work (**STF-Net**), designed to fully leverage the learned representations and their spatiotemporal correlations from the pre-trained echo-video encoder. As shown in Figure 1 **b**, this network can be inserted between the echo-video encoder and various task-specific heads to enhance performance on downstream clinical tasks. The architecture details of the network are presented in Figure 2.

Inefficient use of video representations. As previously discussed, we utilized a vision transformer (ViT) as the video encoder to transform video clips into a sequence of learned video representations, capturing the most expressive information from the raw video. However, due to the joint space-time attention used in the ViT, these learned representations primarily capture global, high-level visual information, often overlooking the spatiotemporal relationships between individual cube embeddings [62]. In natural language processing tasks [73, 74], the relationships between words in a sentence are sequential and one-dimensional, allowing each word to attend to all others. In contrast, the three-dimensional correlations among video patches are essential and should not be simplified through standard mean pooling operations [71, 26], as the relative position of each patch significantly influences the information it conveys. This limitation motivates us to develop a novel approach to maximize the utility of the learned video representations. To address this, we designed a customized spatial-temporal fusion network. The STF-Net takes a sequence of learned video representations as input and outputs a deeply fused representation. To account for the distinct yet complementary relationships between spatial and temporal dimensions, we proposed two separate feature fusion pathways that enable deeper integration of these dimensions.

Joint space-time fusion pathway. A batch of learned video representations, with shape $N \times L_{seq} \times d_{model}$ (where N is the mini-batch size, L_{seq} is the length of video representations, and d_{model} is the feature dimension of echo-video encoder), was first reshaped into a three-dimensional feature map of size $N \times d_{model} \times \bar{t} \times \bar{h} \times \bar{w}$, where

$\overline{t} = \frac{T}{\tau}, \overline{h} = \frac{H}{h}, \overline{w} = \frac{W}{w}$. Note that $L_{seq} = \overline{t} * \overline{h} * \overline{w}$. We then added dense 3D positional encodings [75, 76] to the feature maps before passing them through a ResNet block that performs 3D convolutional operations. This step enhances the 3D relationships within the feature map. The resulting feature map was then flattened while preserving the batch size and feature dimension. Finally, a gated attention mechanism [65] was applied to assign self-attention weights to each flattened representation, generating a compact joint space-time video representation with an output size of $N \times d_{model}$.

Disjoint space-time fusion pathway. In contrast to the joint pathway, we treated the learned representations as a 2D feature map of size $N * \overline{t} \times d_{model} \times \overline{h} \times \overline{w}$, which allows us to learn spatial and temporal relationships in two separate stages. This 2D feature map was processed through a 2D ResNet block [64] for spatial information extraction, followed by a large kernel 2D convolutional operation (with a stride of $(\overline{h}, \overline{w})$) to compress the feature map and capture abstract semantic representations. The resulting output was reshaped into sequential vectors of size $N \times \overline{t} \times d_{model}$. Since downsampling the frames from the original echocardiogram video would result in large temporal intervals, the original frame indices were encoded as 1D positional encodings [73], and adjacent encodings were averaged to reduce the sequence length by half. These averaged encodings were then transformed and adapted through a fully connected layer, producing learnable, sparse positional encodings of size $N \times \overline{t} \times d_{model}$. We then added these sparse positional encodings element-wise to the sequential vectors. Finally, the sequential vectors were aggregated using a gated self-attention mechanism [65]. The output of this pathway also had size $N \times d_{model}$.

The outputs from both pathways were concatenated along the feature dimension, resulting in a batch of fused representations of size $N \times 2 * d_{model}$. This fused representation serves as input for the task-specific head. Our experiments demonstrate that the proposed network is highly effective, yielding significant improvements across a range of downstream tasks.

3.3 Model architectures

We deployed the base version of the vision transformer (ViT) [72] with joint space-time attention [9, 71] as the backbone for the video encoder, consisting of a total of 12 attention layers. Following the principles of masked auto-encoding [51, 26], we used a shallow decoder with 4 layers to recover the raw input video clip. This asymmetric structure enhances the pre-training task, making it more challenging and helping the model learn more informative representations. Additionally, STF-Net is composed of several key components. The Res-block, based on [64], consists of three stacks: convolutional layers, activation functions, normalization, and residual connections. The large-kernel convolution employs the same kernel size and stride, acting as a learnable pooling operation. The gated self-attention mechanism contains three parameter matrices: w $\in \mathbb{R}^{L \times 1}, U \in \mathbb{R}^{L \times d_{model}}, V \in \mathbb{R}^{L \times d_{model}}$, where L is set to 1024. The detailed calculation and derivation were presented here [65]. The fully connected layer consists of two linear transformations interspersed with a ReLU activation function.

3.4 Study approval

This study exclusively used publicly available data.

4 Experiments

In this section, we began by introducing the datasets used to pre-train our echo-video encoder and those employed to evaluate the utility and advantages of the proposed Echo-Vision-FM in §4.1 . Next, we outlined all implementation details and optimization strategies in §4.2. In §4.3, we presented the main experimental results, comparing the performance of our framework to previous state-of-the-art AI models [26, 77, 71] Finally we conducted ablation studies (§4.4) to investigate the pre-training mechanism and analyze the impact of the amount of fine-tuning data on the performance of the pre-trained echo-video encoder.

4.1 Datasets

MIMIC-IV-ECHO. We utilized the MIMIC-IV-ECHO dataset [78] for pre-training our echo-video encoder, which is part of the larger MIMIC-IV database [79]. This dataset comprises 525,328 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, 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 $3 \times 224 \times 224$. 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.

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 (LV_{EF}), left ventricular end-systolic volume (LV_{ESV}), and left ventricular end-diastolic volume (LV_{EDV}), are associated with each video.

CAMUS. The CAMUS dataset [80] consists of 500 fully anonymized clinical echocardiographic exams from the University Hospital of St. Etienne (France), capturing at least one full cardiac cycle in apical four- and two-chamber views, with no specific selection criteria, reflecting routine clinical practice. Half of the patients have a left ventricular ejection fraction (LV_{EF}) below 45%, and 19% of the images are of poor quality. Although these low-quality images are excluded from clinical metric calculations, they are retained for certain training and validation steps. The dataset is divided into 450 exams for training (with corresponding expert manual references) and 50 for testing, all provided in raw/mhd file format. All acquisitions were performed using GE Vivid E95 ultrasound scanners with a GE M5S probe and exported from EchoPAC software. The polar-coordinate B-mode images were consistently re-sampled into a Cartesian grid with a resolution of 0.3 mm along the x-axis and 0.15 mm along the z-axis. Manual annotations of cardiac structures at end-diastole (ED) and end-systole (ES) are included, enabling the calculation of LV_{EF} and other clinical indices. We used the provided codebase [80] to compute LV_{EF}, LV_{ESV}, LV_{EDV} values for individual echocardiogram videos.

TMED. The TMED dataset [81, 82] contains transthoracic echocardiogram (TTE) images collected during routine care at Tufts Medical Center from 2011 to 2020, following ASE guidelines. In practice, a sonographer captures multiple cineloops from various anatomical views using a handheld transducer, from which a single still image is extracted for each video clip, resulting in approximately 50 to 100 images per patient. Routine acquisitions do not include immediate view or diagnostic labels; therefore, expert sonographers or cardiologists manually provided view annotations (e.g., crucial for diagnosing valve diseases like aortic stenosis), while diagnostic severity labels were meticulously extracted from patient records through intensive manual effort. For model evaluation, we used the latest version of the TMED-2 dataset [82], which is fully labeled. It includes 599 studies from 577 unique patients, all of whom have an aortic stenosis (AS) diagnostic label. The available AS severity labels are: "no_as", "mild_as", "mild-to-moderate_AS", "moderate_AS", "severe_AS".

4.2 Implementation details

We utilized the large-scale, publicly available MIMIC-IV-ECHO dataset [79] to train the first stage of our framework in a self-supervised manner, resulting in a generalizable echo-video encoder for vision-related clinical tasks. We set input size $T \times 3 \times H \times W$ as $16 \times 3 \times 224 \times 224$. For the cube embeddings, we configured τ, h, w as $2, 16, 16$, respectively. The feature dimension of learned representation d_{model} is set to 768. For optimization, we employed the AdamW optimizer [83] with $\beta_1 = 0.9, \beta_2 = 0.95$, and a weight decay of 0.05. We used a cosine learning rate decay, with a peak learning rate of 1e-4. During the first 5% of training steps, we applied a warm-up strategy with an initial learning rate of 1e-5. The pre-training process was running on NVIDIA A100 GPUs with mixed precision and Accelerate library [84] for speed optimization. In the fine-tuning stage, we employed a simple multi-layer perceptron (MLP) as the task head to produce the final outputs. For regression tasks, we used mean squared error loss, and for classification tasks, we used cross-entropy loss. The AdamW optimizer [83] with $\beta_1 = 0.9, \beta_2 = 0.98$, and a weight decay of 1e-3, combined with cosine learning rate decay. The fine-tuning process ran on a single 80GB NVIDIA A100 GPU for each task. For consistent input sizes, we resized all frames from the video clip to $3 \times 224 \times 224$ and uniformly sampled 16 frames from the entire video clip. Additionally, we augmented the sampled video data with random flips applied with a probability of 0.5.

4.3 Main Results

In this section, we evaluated the full version of Echo-Vision-FM. The echo-video encoder used here was trained on 100% of the pre-training data with an 85% mask ratio (detailed analyses in §4.4). From a machine learning perspective,

we treated the echo-video encoder as a video feature extractor. By adding a task-specific head on top and performing end-to-end fine-tuning, the model is able to predict key information or assess heart-related functions. To demonstrate the superiority of our method and ensure a fair comparison, we selected several video feature extractors pre-trained in different ways. Specifically, we chose VideoResNet [77] and Vivit [71] as representatives of fully supervised pre-training. Additionally, we included VideoMAE [26] from recent self-supervised video learning frameworks. We reimplemented these three models and loaded the pre-trained weights from their official releases. The task-specific head remains the same across all feature extractors for each downstream task. Furthermore, we inserted our novel STF-Net (see §3.2) between the echo-video encoder and the task-specific head to evaluate its impact. All downstream datasets were divided into training, validation, and test sets according to their official splits. Our comprehensive experimental results (see Table1, 2) showed that our method achieves superior or competitive performance across a variety of labeled datasets from diverse healthcare systems.

Cardiac morphological value estimation. Cardiac morphological value estimation involves measuring and assessing various structural parameters of the heart, often using imaging or direct examination. These measurements provide valuable insights into chamber size, wall thickness, and other relevant features for diagnosing and managing cardiovascular conditions. In this study, we directly estimate left ventricular ejection fraction (LV_{EF}), left ventricular end-systolic volume (LV_{ESV}), and left ventricular end-diastolic volume (LV_{EDV}) from individual echocardiogram videos, framing the task as a univariate regression problem. Our echo-video encoder with STF-Net consistently outperformed the separate echo-video encoder, demonstrating superior performance across all tasks. Specifically, on the Echonet-Dynamic dataset, the echo-video encoder with STF-Net predicted LV_{EF} with a MAE of 3.87% and an r^2 of 0.825. For LV_{ESV}, the MAE was 9.43ml, and for LV_{EDV}, the MAE was 16.04ml, with an r^2 of 0.742. These results represented the best performance compared to other methods. On the CAMUS dataset, the echo-video encoder with STF-Net achieves optimal predictions for LV_{EF} with a MAE of 4.42% and an r^2 of 0.658. For LV_{ESV}, ther^2 was 0.697, and for LV_{EDV}, the MAS was 15.87ml, with an r^2 of 0.682. Notably, VideoResNet performed competitively on certain task metrics, occasionally outperforming other methods, indicating the potential advantages of convolution-based models for some tasks. A complete set of results is presented in Table 1.

Heart function and disease severity diagnosis. Heart function and disease severity diagnosis are essential steps in diagnosing, managing, and predicting the progression of cardiovascular conditions. These assessments combine clinical evaluation, imaging studies, and functional tests. In our experiments, following the work of [17, 85, 80], we evaluated heart function based on a threshold of LV_{EF} in the Echonet-Dynamic and CAMUS datasets, framing this as a binary classification task. Additionally, we diagnosed the severity of aortic stenosis by classifying diagnostic labels from the TMED2 dataset [82], which forms a multi-class classification task. For all tasks, we reported accuracy, AUC (Area Under the ROC Curve), and F1 score. The results demonstrated that the combination of echo-video encoder and STF-Net delivers optimal performance in heart function evaluation. On the Echonet-Dynamic dataset, we achieved an accuracy of 0.905, an AUC of 0.931, and an F1 score of 0.941. For the CAMUS dataset, the performance was also strong, with an accuracy of 0.765, an AUC of 0.862, and an F1 score of 0.789. When diagnosing aortic stenosis on the TMED dataset, the echo-video encoder with STF-Net achieved the best AUC of 0.849 and a suboptimal F1 score of 0.555. Complete results can be found in Table 2.

The performance comparison for morphological value estimation (e.g, LV_{EF}, LV_{ESV}, LV_{EDV}) across all models used is further visualized in Supplementary Figure 1. It is important to note that during training, LV_{ESV} and LV_{EDV} were logarithmically scaled for stability. However, when testing on the held-out set, the predicted values are rescaled back to their raw range to compute the evaluation metrics. Supplementary Figure 2 presents the ROC curves for heart function and disease diagnosis tasks, allowing a direct comparison of model performance. Supplementary Figure 3 demonstrates that Echo-Vision-FM continues to outperform other deep learning models across the same tasks, even when evaluated on different datasets.

4.4 Ablation studies

In this section, we analyzed the effects of different training setups and varying amounts of fine-tuning data on downstream clinical tasks. Understanding the capacity of pre-trained encoders from different pre-training configurations was crucial for guiding future research and maximizing the practical value of these models in medical applications. It is important to note that we do not consider the impact of the proposed STF-Net in this analysis, focusing solely on the performance of the echo-video encoder with simple task-specific prediction heads added on top.

Probing the effect of mask ratio. The mask ratio is one of the most influential hyperparameters in the masked auto-encoding strategy. Recent studies [51, 26] suggested that for visual content such as images and videos, a higher mask ratio (e.g. 70% ~90%) typically resulted in better performance in terms of learned representations. In line with

Table 1: Results of cardiac morphological value estimation: For each individual regression task, we reported the MAE and r^2 are reported as metric. Optimal results were highlighted in bold, while suboptimal results are underlined. The values in parentheses represent the 95% confidence interval. All results were reported on the held-out test set..

Datasets		EchoNet-Dynamic						CAMUS					
Methods		LV_{EF}		LV_{ESV}		LV_{EDV}		LV_{EF}		LV_{ESV}		LV_{EDV}	
		mae	r^2	mae	r^2	mae	r^2	mae	r^2	mae	r^2	mae	r^2
Supervised	VideoResNet	5.44% (±0.02%)	0.632 (±0.013)	10.64ml (±0.46)	0.754 (±0.029)	17.34ml (±0.56)	0.685 (±0.022)	6.21% (±0.10%)	0.400 (±0.033)	**12.00ml** (±0.20)	0.661 (±0.019)	19.40ml (±0.40)	0.578 (±0.019)
	Vivit	6.68% (±0.03%)	0.451 (±0.014)	16.21ml (±0.80)	0.350 (±0.049)	23.66ml (±0.91)	0.365 (±0.045)	6.95% (±0.04%)	0.007 (±0.014)	17.06ml (±0.12)	0.209 (±0.010)	21.00ml (±0.70)	0.290 (±0.012)
Natural SSVL	VideoMAE	5.05% (±0.03%)	0.699 (±0.013)	12.77ml (±0.59)	0.651 (±0.041)	20.99ml (±0.87)	0.555 (±0.024)	5.73% (±0.04%)	0.339 (±0.021)	15.52ml (±0.15)	0.337 (±0.011)	19.00ml (±0.10)	0.423 (±0.006)
Medical SSVL	Ours (w/o STF)	4.56% (±0.04%)	0.757 (±0.011)	9.81ml (±0.49)	**0.785** (±0.023)	21.82ml (±0.73)	0.525 (±0.036)	5.53% (±0.11%)	0.609 (±0.032)	17.60ml (±0.30)	0.388 (±0.036)	20.40ml (±0.50)	0.549 (±0.036)
	Ours (w STF)	**3.87%** (±0.01%)	**0.825** (±0.016)	**9.43ml** (±0.39)	0.782 (±0.024)	**16.04ml** (±0.55)	**0.742** (±0.011)	**4.42%** (±0.17%)	**0.658** (±0.034)	15.14ml (±0.32)	**0.697** (±0.042)	**15.87ml** (±0.43)	**0.682** (±0.036)

Table 2: Results of heart function and disease severity diagnosis: for each individual classification task accuracy, AUC and F1 score were reported as metric. Bold font and underline represent optimal and suboptimal results, respectively. Following [85, 80] thresholds of LV_{EF} were set to be 0.5 and 0.45 for Echonet-Dynamic and CAMUS respectively. All results were reported on held-out test set.

Datasets		EchoNet-Dynamic			CAMUS			TMED		
Methods		LV_{EF}			LV_{EF}			Aortic Stenosis		
		Acc	AUC	F1 score	Acc	AUC	F1 score	Acc	AUC	F1 score
Supervised	VideoResNet	0.875 (±0.009)	0.898 (±0.006)	0.922 (±0.004)	0.661 (±0.008)	0.783 (±0.009)	0.708 (±0.010)	0.616 (±0.003)	0.828 (±0.003)	0.473 (±0.004)
	Vivit	0.837 (±0.005)	0.854 (±0.009)	0.901 (±0.011)	0.573 (±0.007)	0.622 (±0.009)	0.695 (±0.006)	0.571 (±0.009)	0.823 (±0.005)	0.469 (±0.007)
Natural SSVL	VideoMAE	0.879 (±0.009)	0.916 (±0.008)	0.923 (±0.002)	0.559 (±0.008)	0.567 (±0.010)	0.710 (±0.007)	0.632 (±0.012)	0.804 (±0.008)	0.508 (±0.011)
Medical SSVL	Ours (w/o STF)	0.878 (±0.009)	0.917 (±0.009)	0.924 (±0.006)	0.708 (±0.014)	0.860 (±0.008)	0.777 (±0.011)	**0.651** (±0.008)	0.828 (±0.009)	**0.563** (±0.013)
	Ours (w STF)	**0.905** (±0.007)	**0.931** (±0.007)	**0.941** (±0.005)	**0.765** (±0.015)	**0.862** (±0.013)	**0.789** (±0.014)	0.620 (±0.011)	**0.849** (±0.007)	0.555 (±0.019)

this, we explored a range of mask ratios from 50% ~90%, training the video encoder on the full pre-training dataset. We then evaluated the resulting echo-video encoders on the widely used Echonet-Dynamic dataset [17], specifically for estimating the key morphological value LV_{EF}. The results are presented in Supplementary Figure 4 and Table 3. Our first observation was that a mask ratio of 85% mask ratio led to the best performance. Additionally, while increasing the mask ratio generally produceed more expressive representations, excessively high mask ratios (e.g., above 85%) could negatively impact the model's transferability.

Table 3: Results for varying mask ratios during pre-training: The echo-video encoders were pre-trained using the full dataset and subsequently evaluated on the EchoNet-Dynamic dataset for the LV_{EF} regression task.

mask ratio	LV_{EF}			
	mae	mse	rmse	r^2
50%	5.05% (±0.02%)	0.005 (±0.001)	0.068 (±0.003)	0.690 (±0.010)
70%	4.68% (±0.03%)	0,004 (±0.000)	0.062 (±0.003)	0.739 (±0.009)
80%	4.61% (±0.03%)	0.004 (±0.001)	0.061 (±0.004)	0.749 (±0.010)
85%	**4.56%** (±0.04%)	**0.003** (±0.001)	**0.060** (±0.004)	**0.757** (±0.016)
90%	4.74% (±0.02%)	0.003 (±0.001)	0.062 (±0.003)	0.741 (±0.013)

Table 4: Results with varying data ratios for pre-training: In this experiment, 85% mask ratio was used during pre-training, and the resulting echo-video encoders were evaluated on the EchoNet-Dynamic dataset for both LV_{EF} regression and classification tasks. Additionally, we conducted both fine-tuning and linear probing to assess the effectiveness of the learned representations.

data ratio	LV_{EF} (reg ft)				LV_{EF} (reg lp)				LV_{EF} (cls ft)			LV_{EF} (cls lp)		
	mae	mse	rmse	r^2	mae	mse	rmse	r^2	Acc	AUC	F1 score	Acc	AUC	F1 score
25%	5.12% (±0.03%)	0.005 (±0.000)	0.068 (±0.001)	0.686 (±0.019)	8.42% (±0.02%)	0.012 (±0.003)	0.108 (±0.013)	0.220 (±0.014)	**0.882** (±0.001)	0.919 (±0.008)	**0.925** (±0.004)	0.816 (±0.009)	0.792 (±0.010)	0.887 (±0.007)
50%	5.04% (±0.08%)	0.005 (±0.001)	0.068 (±0.009)	0.693 (±0.028)	7.99% (±0.04%)	0.011 (±0.005)	0.105 (±0.006)	0.262 (±0.010)	0.876 (±0.005)	**0.927** (±0.003)	0.920 (±0.008)	0.817 (±0.011)	**0.811** (±0.006)	0.887 (±0.001)
75%	4.92% (±0.04%)	0.004 (±0.000)	0.069 (±0.001)	0.705 (±0.014)	8.04% (±0.05%)	0.011 (±0.002)	0.106 (±0.004)	0.246 (±0.020)	0.868 (±0.005)	0.909 (±0.004)	0.915 (±0.007)	0.807 (±0.009)	0.792 (±0.008)	0.884 (±0.002)
100%	**4.56%** (±0.04%)	**0.003** (±0.001)	**0.060** (±0.004)	**0.757** (±0.016)	**7.91%** (±0.02%)	**0.010** (±0.004)	**0.104** (±0.003)	**0.268** (±0.009)	0.878 (±0.008)	0.917 (±0.009)	0.924 (±0.006)	**0.818** (±0.001)	0.802 (±0.004)	**0.888** (±0.001)

Table 5: Results on different fine-tuning data ratio: regression and classification of LV_{EF} on Echonet-Dynamic dataset.

LV_{EF} (reg)						LV_{EF} (cls)								
1%		10%		100%		1%			10%			100%		
mae	r2	mae	r2	mae	r2	Acc	AUC	F1 score	Acc	AUC	F1 score	Acc	AUC	F1 score
17.62% (±0.29%)	-2.231 (±0.128)	8.51% (±0.19%)	0.178 (±0.29)	4.56% ±0.04%	0.757 (±0.011)	0.748 (±0.010)	0.675 (±0.014)	0.844 (±0.007)	0.832 (±0.012)	0.857 (±0.016)	0.892 (±0.008)	0.878 (±0.009)	0.917 (±0.009)	0.924 (±0.006)

Data efficiency for pre-training. In this experiment, we assessed the impact of varying amounts of pre-training data (i.e., 25%, 50%, 75%, 100%) on the performance of the echo-video encoder. After pre-training, we evaluate the encoder's ability to predict LV_{EF} and assess heart function [86] by classifying LV_{EF} with a threshold of 0.50, using both end-to-end fine-tuning and linear probing. The experimental results are presented in Table4. We observed that as the amount of pre-training data increases, the performance consistently improved for both fine-tuning and linear probing, where the feature extractor was frozen.

Low-data regime. Following recent medical self-supervised learning studies [87, 88], we evaluated the impact of varying downstream data amounts on fine-tuning the model to further analyze the utility of the pre-trained echo-video encoder. Using the Echonet-Dynamic dataset, we performed regression and classification tasks on LV_{EF} tasks, with fine-tuning data ratios set at 1%, 10%, 100%. The complete results are presented in Table 5. For the regression task, performance significantly dropped with low data usage (e.g., MAE of 17.62% at 1%), while increased data availability led to substantial improvements (r^2 increased from 0.178 when using 10% to 0.757 when using 100%). For the classification task, performance similarly improved as more data was used. However, unlike regression, the performance gained become more stable as data quantity increased, indicating that classification is less sensitive to data size compared to regression tasks.

5 Discussion

This study demonstrated that large-scale video datasets of echocardiogram studies can serve as a foundation for training and deploying medical video models. Our echocardiogram video vision foundation model, along with the designed two-stage 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 learning and a unique masking strategy yielded strong and transferable learned video representations. Through the use of SSVL and masked auto-encoding, our model outperformed existing state-of-the-art baselines, which often relied on fully supervised learning. Additionally, the proposed STF-Net, designed to leverage the spatiotemporal correlations within video representations, significantly enhanced performance across nearly all assessments of heart function, diseases diagnosis and morphological value predictions.

Echo-Vision-FM 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 [86, 23, 85]. Notably, Echo-Vision-FM outperformed previous models, achieving optimal or competitive results. For instance, it attained a mean absolute error (MAE) of 3.87% in external validation of LV_{EF} value prediction from Echonet-Dynamic, while the most recent video-based LV_{EF} AI model [85] achieved an MAE of 4.34% and a multi-modal AI model [86] reported an MAE of 7.1%. For the

clinically significant LV_{EF} threshold of 50%, Echo-Vision-FM surpassed previous AI models, achieving an AUC of 0.931 compared to the 0.89–0.90 achieved by others [86].

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 [89]. 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 [90, 91, 92]. 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 echo-video 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 believed that this off-the-shelf 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 [86, 93]. We noted that many echocardiogram-related models rely on 2D image frames extracted from the entire echocardiogram [6, 86, 23] rather than directly modeling 3D video clips, which contains 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 echo-video encoder, we developed a novel, simplified spatial-temporal fusion network (STF-Net). This network can be integrated with the echo-video encoder and fine-tuned during the second stage. For standard ViT encoders using the original self-attention mechanism [73, 72], adding a learnable classification token [72] and pooling the last-layer token embeddings are usually 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-Vision-FM, there are several limitations that warrant discussion. First, 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. Second, 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-Vision-FM 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, Echo-Vision-FM offers a promising foundation for developing more efficient, scalable, and accurate diagnostic tools in medicine.

6 Conclusion

In this study, we developed Echo-Vision-FM, a pioneering echocardiogram video foundation model that leverages large-scale, publicly available datasets for training and deployment. Our approach demonstrated significant advancements in the prediction and assessment of cardiac function through a self-supervised video learning framework, highlighting the model's robust transferability and adaptability across diverse clinical settings. By integrating

pre-trained echo-video encoder with a novel STF-Net, we achieved superior performance in various cardiac tasks compared to existing state-of-the-art models. The findings suggested that utilizing a complete echocardiogram video rather than isolated frames provides richer information, leading to more accurate predictions of critical cardiac parameters. Echo-Vision-FM represented 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`. CAMUS dataset is available at `https://www.creatis.insa-lyon.fr/Challenge/camus/`. TMED dataset is available at `https://tmed.cs.tufts.edu/tmed_v2.html`.

Code Availability Statement The relevant code are available at: `https://github.com/ZiyangZhang0511/Echo-Vison-FM.git`.

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

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

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

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

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

[55] 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.

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

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

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

[66] Zelun Wang and Jyh-Charn Liu. Translating math formula images to latex sequences using deep neural networks with sequence-level training, 2019.

[67] 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.

[68] 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.

[69] 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.

[70] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

[71] 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.

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

[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] 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.

[78] 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*.

[79] 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.

[80] 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.

[81] Zhe Huang, Gary Long, Benjamin Wessler, and Michael C. Hughes. A new semi-supervised learning benchmark for classifying view and diagnosing aortic stenosis from echocardiograms. In *Proceedings of the 6th Machine Learning for Healthcare Conference (MLHC)*, 2021.

[82] Zhe Huang, Gary Long, Benjamin Wessler, and Michael C Hughes. Tmed 2: a dataset for semi-supervised classification of echocardiograms. In *In DataPerf: Benchmarking Data for Data-Centric AI Workshop*, 2022.

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

[84] 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.

[85] 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.

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

[87] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

[88] Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3942–3951, October 2021.

[89] 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.

[90] 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.

[91] 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.

[92] 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.

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

Supplementary

1 Additional results

The performance comparison for morphological value estimation (e.g, LV_{EF}, LV_{ESV}, LV_{EDV}) across all models used is visualized in Figure 1. It is important to note that during training, LV_{ESV} and LV_{EDV} were logarithmically scaled for stability. However, when testing on the held-out set, the predicted values are rescaled back to their raw range to compute the evaluation metrics. Figure 2 presents the ROC curves for heart function and disease diagnosis tasks, allowing a direct comparison of model performance. Figure 3 demonstrates that Echo-Vision-FM continues to outperform other deep learning models across the same tasks, even when evaluated on different datasets.

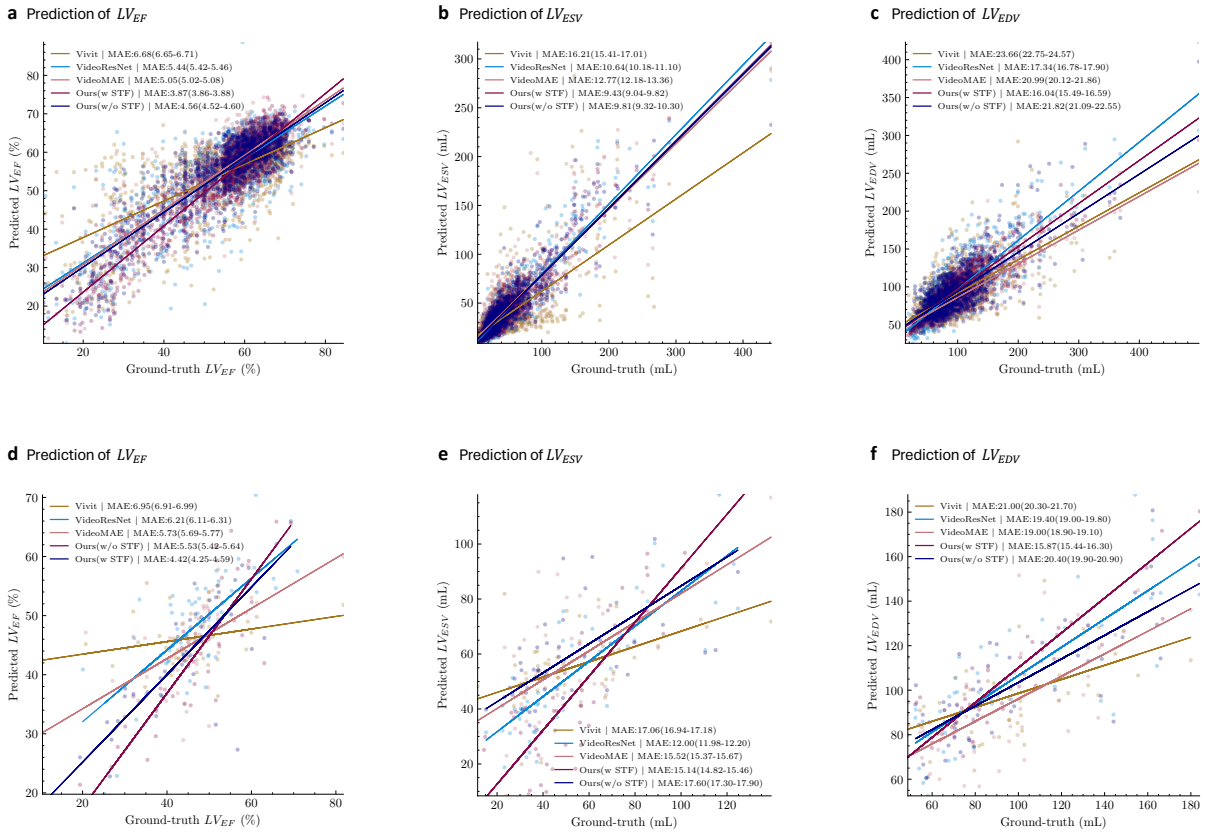

Figure 1: Comparison of all models on cardiac morphological value estimation. **a**, **b**, **c**: Scatter plots of predicted versus actual values for LV_{EF}, LV_{ESV}, and LV_{EDV} on the test set from Stanford Healthcare (Echonet-Dynamic). **d**, **e**, **f**: Scatter plots of predicted versus actual values for LV_{EF}, LV_{ESV}, LV_{EDV} on the test set from the University Hospital of St. Etienne (CAMUS).

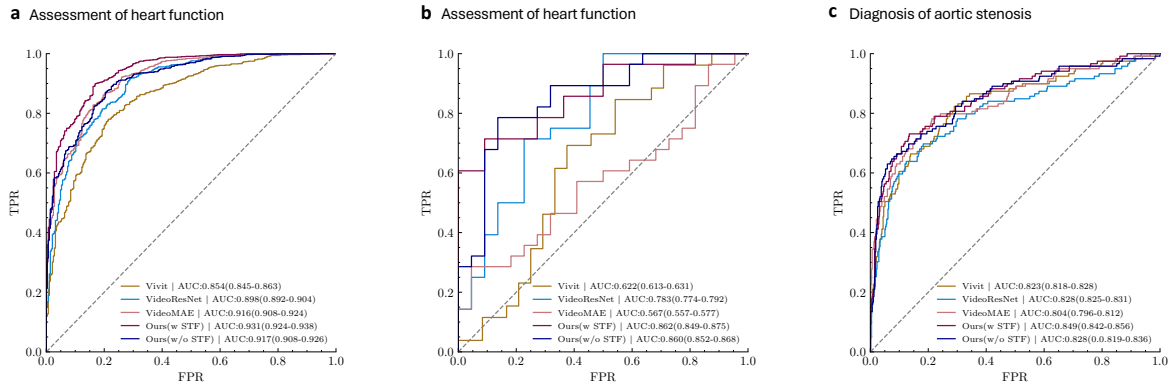

Figure 2: Comparison of all models on heart function and disease diagnosis. **a**, **b**: AUC performance for heart function diagnosis (classifying LV_{EF}) on held-out test dataset from Stanford Healthcare (Echonet-dynamic), University Hospital of St. Etienne (CAMUS). **c**: AUC performance for diagnosis of aortic stenosis (e.g. "no_as", "mild_as", "mild-to-moderate_AS", "moderate_AS", "severe_AS") on the Tufts Medical Center (TMED) dataset.

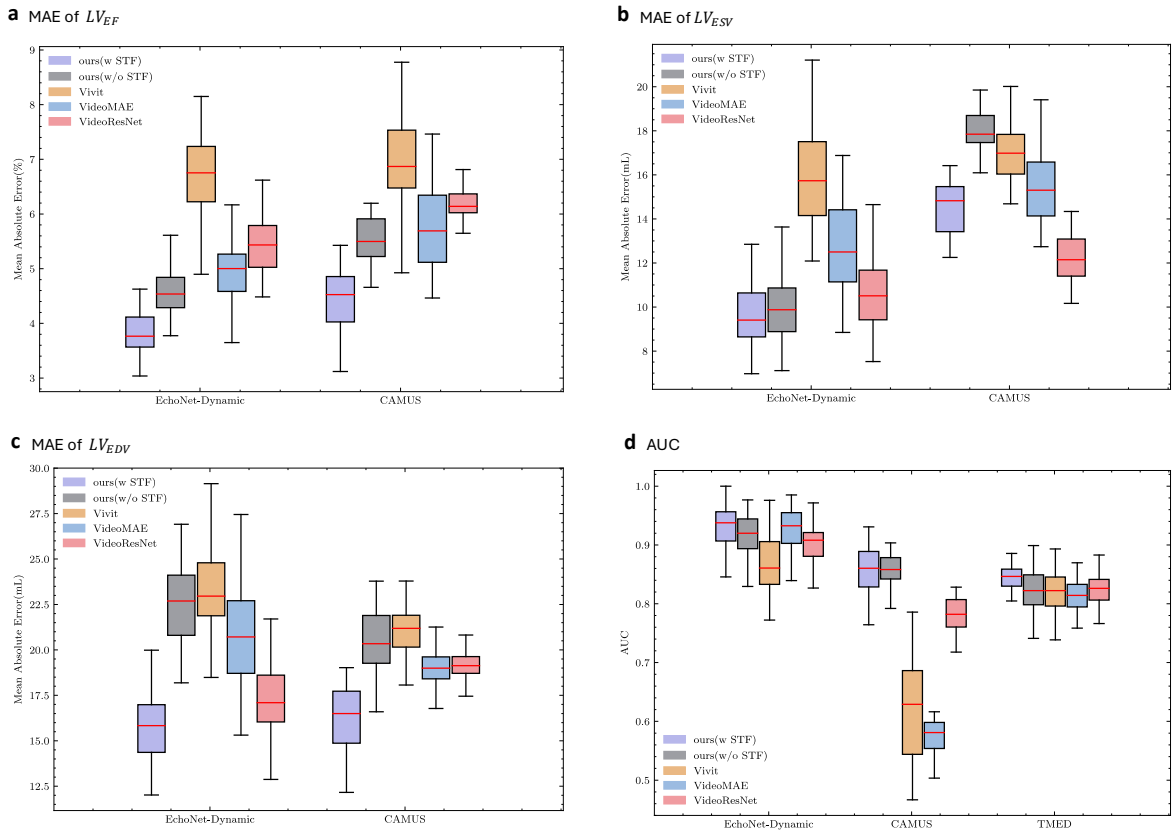

Figure 3: Box-plots of different clinical tasks. The center lines represent the median, and the boxes extend from the first to the third quartile.

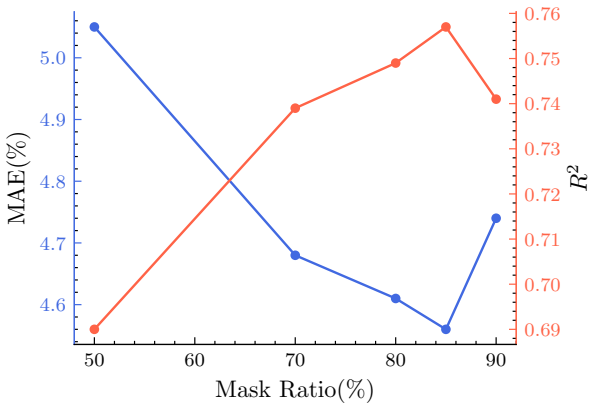

Figure 4: A double y-axis line chart illustrating the LV_{EF} regression results on the EchoNet-Dynamic held-out test set. The blue line represents the MAE, and the red line represents the r^2 values across different mask ratios. The chart demonstrates that the model achieved the best performance at a mask ratio of 85%.