Comprehensive echocardiogram evaluation with view primed vision language AI

https://doi.org/10.1038/s41586-025-09850-x

Received: 12 October 2024

Accepted: 3 November 2025

Published online: 11 November 2025

Open access

Check for updates

Milos Vukadinovic[1,2], I-Min Chiu[1,3], Xiu Tang[4], Neal Yuan[5,6], Tien-Yu Chen[7], Paul Cheng[4], Debiao Li[8], Susan Cheng[1], Bryan He[9,11] ✉ & David Ouyang[1,10,11] ✉

Echocardiography is the most widely used cardiac imaging modality, capturing ultrasound video data to assess cardiac structure and function[1]. Artificial intelligence (AI) in echocardiography has the potential to streamline manual tasks and improve reproducibility and precision[2]. However, most echocardiography AI models are single-view, single-task systems that do not synthesize complementary information from multiple views captured during a full examination[3,4], and thus lead to limited performance and scope of applications. To address this problem, we introduce EchoPrime, a multi-view, view-informed, video-based vision–language foundation model trained on over 12 million video–report pairs. EchoPrime uses contrastive learning to train a unified embedding model for all standard views in a comprehensive echocardiogram study with representation of both rare and common diseases and diagnoses. EchoPrime then utilizes view classification and a view-informed anatomical attention module to weight video-specific embeddings that accurately map the relationship between echocardiographic views and anatomical structures. With retrieval-augmented interpretation, EchoPrime integrates information from all echocardiogram videos in a comprehensive study and performs holistic clinical interpretation. In datasets from five international independent health-care systems, EchoPrime achieves state-of-the-art performance on 23 diverse benchmarks of cardiac form and function, surpassing the performance of both task-specific approaches and previous foundation models. Following rigorous clinical evaluation, EchoPrime can assist physicians in the automated preliminary assessment of comprehensive echocardiography.

In recent years, medical AI has progressed substantially, driven by the fast improvements of deep learning methods and the use of larger and larger medical datasets. AI has matched or surpassed the accuracy of clinical experts in various applications[5], such as skin cancer classification[6] and breast mammogram lesion detection[7]. However, because these models are tailored to specific tasks through supervised learning, they lack the ability to synthesize information in a holistic manner, unlike clinical experts that integrate multiple data points for a comprehensive assessment. Moreover, with thousands of possible diagnoses and diseases, it is impractical to train separate models for every individual medical task. This limitation gave rise to foundation models[8–16], task-agnostic models pre-trained on large datasets that demonstrate robust performance in various downstream tasks. Foundation models have already been utilized across various medical subfields, including pathology[17], drug repurposing[18], chest X-ray[19] and retinal imaging[20].

Echocardiography, or cardiac ultrasound, is the most common form of cardiac imaging and benefits from high volume, low cost, portability and lack of ionizing radiation. With the highest temporal resolution across all imaging modalities, echocardiography videos capture changes in heart motion and structure associated with cardiomyopathy, valvular disorders, tamponade and arrhythmias. Progress towards a foundation model for echocardiography was made with EchoCLIP[3], which was trained on over 1 million echocardiogram videos and demonstrated good performance across a diverse range of benchmarks for cardiac image interpretation. However, EchoCLIP utilizes only a static image encoder from a single echocardiogram view, rather than incorporating available dynamic videos from a comprehensive ultrasound examination. As a result, it might miss key temporal and functional insights that are vital for echocardiographic analysis.

In this work, we introduce EchoPrime, a video-based foundation model for echocardiography, trained with contrastive learning on more than 12 million videos paired with expert interpretations and designed to synthesize data from multiple videos to deliver

[1]Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [2]Department of Bioengineering, University of California Los Angeles, Los Angeles, CA, USA. [3]Department of Emergency Medicine, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan. [4]Division of Cardiology, Department of Medicine, Stanford University, Palo Alto, CA, USA. [5]Department of Medicine, University of California, San Francisco, CA, USA. [6]Division of Cardiology, San Francisco Veterans Affairs Medical Center, San Francisco, CA, USA. [7]Division of Cardiology, Department of Medicine, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung, Taiwan. [8]Biomedical Imaging Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [9]Department of Computer Science, Stanford University, Stanford, CA, USA. [10]Division of Research, Kaiser Permanente Northern California, Pleasanton, CA, USA. [11]These authors contributed equally: Bryan He, David Ouyang. ✉e-mail: bryanhe@cs.stanford.edu; david.ouyang@kp.org

Table 1 | Clinical characteristics of the study cohorts

Characteristic	CSMC	SHC	MIMIC	CGMH	KP
Videos (*n*)	12,124,168	91,746	75,768	24,724	201,752
Studies (*n*)	275,442	1,792	2,431	1,188	4,891
Patients (*n*)	108,913	1,779	1,767	1,188	4,891
Age (mean (s.d.))	66.31 (16.7)	60.42 (17.3)	67.67 (13.5)	65.87 (16.4)	67.28 (14.6)
Female (*n* (%))	117,324 (42.6)	843 (47.0)	909 (51.4)	583 (49.1)	2,392 (48.0)
Race (*n* (%))					
Non-Hispanic white	166,091 (60.3)	900 (50.2)	1,234 (69.8)	0 (0.0)	2,990 (61.1)
Black	35,624 (12.9)	76 (4.2)	343 (19.4)	0 (0.0)	459 (9.4)
Hispanic	27,194 (9.9)	241 (13.5)	68 (3.8)	0 (0.0)	50 (1.0)
Asian	20,299 (7.4)	301 (16.8)	45 (2.5)	1,188 (100.00)	786 (16.1)
Other	19,832 (7.2)	161 (9.0)	42 (2.4)	0 (0.0)	496 (10.1)
Unknown	5,321 (1.9)	85 (4.7)	35 (2.0)	0 (0.0)	110 (2.2)
Pacific Islander	984 (0.4)	28 (1.6)	0 (0.0)	0 (0.0)	0
Conditions (*n* (%))					
Hypertension	112,919 (41.0)	1,055 (59.9)	1,510 (85.5)	–	2,142 (43.8)
Heart failure	93,875 (34.1)	746 (41.6)	996 (56.4)	–	1,015 (20.8)
Atrial fibrillation	60,448 (21.9)	555 (30.0)	510 (28.9)	–	937 (19.2)
Chronic kidney disease	50,302 (18.3)	580 (32.4)	755 (42.7)	–	820 (16.8)
Diabetes mellitus	46,656 (16.9)	168 (9.3)	486 (27.5)	–	1,038 (21.2)
Pulmonary artery disease	28,903 (10.5)	41 (2.3)	141 (8.0)	–	250 (5.1)
Cerebrovascular accident	18,442 (6.7)	104 (5.8)	160 (9.1)	–	412 (8.4)
Myocardial infarction	18,283 (6.6)	296 (16.5)	507 (28.7)	–	460 (9.4)

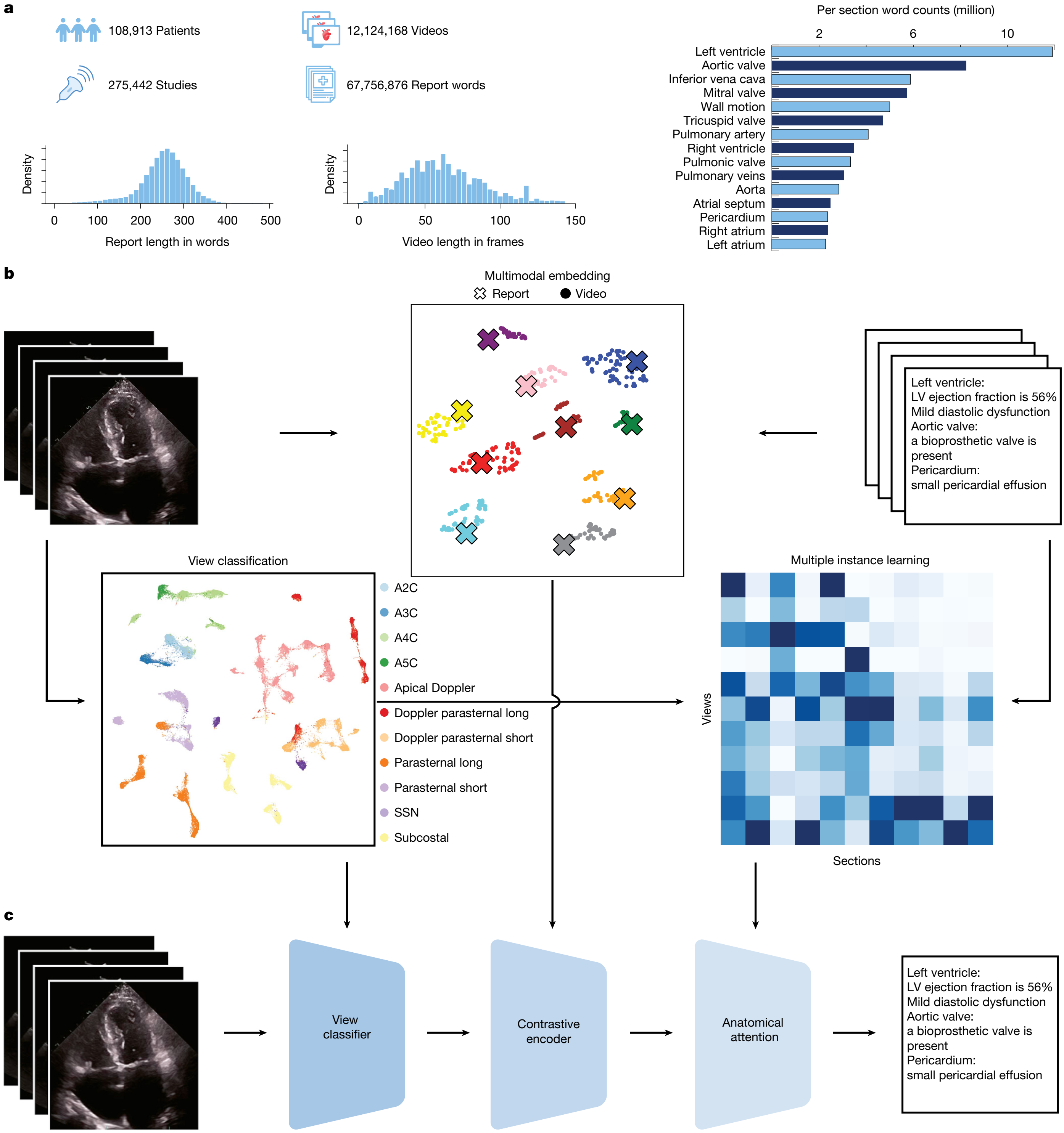
comprehensive interpretations. We tested the performance of the model on datasets across five international health-care systems on an extensive benchmark of multi-modal retrieval metrics, clinical echocardiography interpretation tasks covering all cardiac structures, and transfer learning tasks related to cardiac pathophysiology. We found that EchoPrime consistently outperforms other medical foundation models (BioMedCLIP and EchoCLIP) and either matches or exceeds the performance of task-specific ultrasound models. In additional analyses, we found that EchoPrime prioritizes views similar to clinical experts when assessing many echocardiogram videos through multiple instance attention. EchoPrime is the largest existing echocardiography AI model trained on over ten times the data of previous models and natively provides multi-view, multi-task and multi-video assessments. To advance AI in medicine research, we have publicly released code, weights and a demo.

EchoPrime is a video-based vision–language model trained with 12,124,168 echocardiography videos and paired text reports from 275,442 studies across 108,913 patients at Cedars-Sinai Medical Center (CSMC; Table 1). EchoPrime consists of multiple modules integral for the interpretation of echocardiography, including a video encoder, text encoder, view classifier and anatomical attention module (Fig. 1). The video and text encoders are trained contrastively on sampled echocardiogram video clips and corresponding cardiologist report texts to learn a joint video–text representation space. A view classification model was trained on 77,426 sonographer-labelled videos to classify B-mode and colour Doppler videos into 58 standard echocardiographic views and utilized by an anatomical attention module to determine the relative importance of each echocardiogram video for interpretation tasks using multiple instance learning. During inference, EchoPrime provides a comprehensive interpretation of an echo study by assigning views to each video, mapping the videos into a contrastive joint video–text latent space, and finally retrieving study interpretations guided by the anatomical attention module.

Automated echocardiogram interpretation

EchoPrime can simultaneously interpret a wide range of cardiac features and diagnoses that represent a wide range of pathophysiology seen over a decade of echocardiography at a large tertiary care centre (Supplementary Video 1). To obtain echocardiogram examination interpretations, we used retrieval augmented interpretation, a method similar to retrieval augmented generation[21]. Retrieval augmented interpretation works by retrieving the historical echocardiogram reports that best match input echocardiogram videos, and then filters the information from these reports into final output interpretations based on anatomical attention (see Methods for details). This approach allows a single foundation model to be applied directly to all echocardiography interpretation tasks, enabling the prediction of hundreds of pathologies across 15 sections of over 1,000 different statements that cardiologists frequently make in interpreting echocardiograms.

We evaluated EchoPrime on a diverse set of benchmarks for cardiac structure and function across datasets from five different institutions (Table 2). On echocardiographic data from all five geographically diverse hospitals, EchoPrime outperformed previously developed foundation models for both general biomedical imaging and echocardiography[3,4,22–24] (Fig. 2 and Supplementary Tables 1–5), and without additional fine-tuning or training, EchoPrime outperforms or matches fully supervised models for single-task prediction[25–30] (Extended Data Table 1). EchoPrime had a mean area under the curve (AUC) of 0.92 on CSMC, 0.89 on Stanford Healthcare (SHC), 0.85 on Beth Israel Deaconess Medical Center (MIMIC), 0.86 on Chang Gung Memorial Hospital (CGMH) and 0.88 on Kaiser Permanente (KP) health-care system across 17 classification tasks and 11 regression tasks (Table 2, Extended Data Table 2 and Supplementary Fig. 1). On a held-out internal validation and external validation cohorts from four international hospitals, EchoPrime outperformed both previously developed foundation models and task-specific echocardiography AI models. Without additional



Left ventricle:
LV ejection fraction is 56%
Mild diastolic dysfunction
Aortic valve:
a bioprosthetic valve is present
Pericardium:
small pericardial effusion

Left ventricle:
LV ejection fraction is 56%
Mild diastolic dysfunction
Aortic valve:
a bioprosthetic valve is present
Pericardium:
small pericardial effusion

Fig. 1 | EchoPrime overview. a, Training dataset characteristics. The schematics were adapted from Flaticon by FreePik (https://www.flaticon.com). **b**, Video and text encoders are trained to map videos and text into a joint latent space. The view classifier and multiple instance learning are trained separately to learn view weighting based on the anatomical section. LV, left ventricle. SSN, suprasternal notch. **c**, Inference pipeline. When given a comprehensive echocardiogram study, EchoPrime determines the view for each video, uses contrastive encoder on the joint video–text latent space for retrieval augmented interpretation, and weights the interpretation of each video by anatomical attention for a study-level assessment of each cardiac structure.

fine-tuning or training, EchoPrime outperforms or matches fully supervised models for a single-task prediction[25–30] including a 2% improvement over EchoNet-Dynamic[25] in R^2 score for left ventricular ejection fraction, 2% improvement in the area under the receiver operating characteristic curve (AUC) over Echonet-TR[26] for assessing tricuspid regurgitation, 4% increase over EchoNet-MR[27] for assessing mitral regurgitation, and similar to improved performances for pericardial effusion[28] and aortic stenosis[29]. Compared with other echocardiography AI models, EchoPrime has superior performance (Extended Data Table 3 and Supplementary Tables 1–5), particularly driven by differences in low-quality and low-view-count images when informed by anatomical attention (Extended Data Table 4).

Table 2 | EchoPrime echocardiography interpretation performance across datasets from five different medical centres

Task	CSMC	SHC	MIMIC	CGMH	KP
Left ventricle ejection fraction (R^2)	0.83 (0.81–0.85)	0.79 (0.76–0.82)	0.74 (0.66–0.80)	0.51 (0.44–0.57)	0.66 (0.62–0.71)
Left ventricle ejection fraction (MAE)	4.79 (4.60–4.97)	4.14 (3.94–4.34)	6.21 (5.54–6.92)	6.46 (6.19–6.74)	4.51 (4.30–4.72)
Pacemaker	0.85 (0.82–0.88)	0.84 (0.81–0.87)	0.86 (0.80–0.92)	0.78 (0.63–0.92)	0.84 (0.81–0.87)
Right ventricle systolic function	0.93 (0.91–0.95)	0.94 (0.88–1.00)	0.98 (0.97–0.99)	0.90 (0.79–0.98)	0.92 (0.88–0.95)
Right ventricle dilation	0.89 (0.83–0.95)	0.85 (0.80–0.90)	–	0.87 (0.69–0.97)	0.78 (0.73–0.83)
Left atrial dilation	0.91 (0.86–0.96)	0.73 (0.70–0.75)	0.72 (0.68–0.75)	0.65 (0.62–0.67)	0.77 (0.75–0.79)
Right atrial dilation	0.90 (0.82–0.97)	0.77 (0.74–0.81)	0.65 (0.61–0.69)	0.75 (0.67–0.84)	0.82 (0.79–0.84)
Mitraclip	0.99 (0.99–1.00)	0.98 (0.94–1.00)	0.99 (0.99–1.00)	–	1.00 (1.00–1.00)
Mitral annular calcification	0.96 (0.95–0.97)	0.96 (0.94–0.98)	0.88 (0.85–0.91)	0.88 (0.75–0.96)	0.89 (0.86–0.91)
Mitral stenosis	0.96 (0.91–0.99)	0.92 (0.86–0.98)	0.96 (0.95–0.98)	0.79 (0.59–0.98)	0.88 (0.78–0.95)
Mitral regurgitation	0.92 (0.91–0.94)	0.91 (0.89–0.93)	0.81 (0.79–0.83)	0.86 (0.82–0.89)	0.89 (0.86–0.91)
TAVR	1.00 (0.99–1.00)	0.97 (0.90–1.00)	0.96 (0.94–0.98)	1.00 (1.00–1.00)	0.99 (0.98–0.99)
Bicuspid aortic valve	0.83 (0.67–0.96)	0.82 (0.73–0.90)	0.65 (0.52–0.78)	0.67 (0.45–1.00)	0.80 (0.72–0.88)
Aortic stenosis	0.98 (0.96–0.99)	0.96 (0.92–0.99)	0.88 (0.84–0.91)	0.99 (0.98–1.00)	0.97 (0.96–0.98)
Aortic regurgitation	0.88 (0.83–0.93)	0.89 (0.80–0.97)	0.81 (0.75–0.86)	0.78 (0.70–0.84)	0.87 (0.83–0.92)
Tricuspid regurgitation	0.95 (0.93–0.97)	0.88 (0.86–0.91)	0.84 (0.82–0.86)	0.93 (0.90–0.95)	0.91 (0.89–0.93)
Pericardial effusion	0.98 (0.95–0.99)	0.89 (0.77–0.98)	0.87 (0.83–0.92)	0.96 (0.91–1.00)	1.00 (0.99–1.00)
Aortic root dilation	0.91 (0.83–0.98)	0.97 (0.94–.99)	0.98 (0.96–0.99)	0.98 (0.97–1.00)	0.81 (0.78–0.85)
Dilated IVC	0.84 (0.82–0.86)	0.86 (0.81–0.91)	0.59 (0.49–0.69)	0.98 (0.94–1.00)	0.80 (0.77–0.82)
Diastolic dysfunction	0.91 (0.88–0.93)	0.86 (0.81–0.90)	0.93 (0.86–0.98)	–	0.84 (0.81–0.87)
PA pressure (R^2)	0.43 (0.39–0.47)	0.36 (0.30–0.42)	–	–	0.31 (0.27–0.35)
PA pressure (MAE)	7.30 (6.95–7.67)	7.97 (7.35–8.66)	–	–	7.39 (7.08–7.70)

For classification tasks, we report the area under the receiver operating characteristic curve, and regression tasks are evaluated using R^2 and MAE. AV, aortic valve; PA, pulmonary artery. The values in parentheses show 95% confidence intervals.

Our model performed favourably when compared with other foundation models such as BioMedCLIP[4], a general foundation model for biomedicine, and EchoCLIP[3], a previous foundation model for echocardiography (Fig. 2). EchoPrime demonstrated significant improvements over previous models in predicting features related to the motion of cardiac structures. Specifically, in estimating left ventricular ejection fraction, EchoPrime achieved a mean absolute error (MAE) of 4.79 on the internal CSMC dataset, compared with MAEs of 26.93 and 7.00 for BioMedCLIP and EchoCLIP, respectively. On external validation cohorts, EchoPrime had MAE ranging from 4.14 to 6.46, whereas BioMedCLIP and EchoCLIP had MAEs ranging from 23.00 to 32.64 and 6.23 to 10.37, respectively. In detecting aortic regurgitation, EchoPrime achieved an AUC of 0.88 versus 0.54 and 0.68 for BioMedCLIP and EchoCLIP in the internal dataset, and 0.78–0.89 versus 0.53–0.58 and 0.61–0.67 for BioMedCLIP and EchoCLIP, respectively, in the external datasets (full results are in Table 2). Our results show that EchoPrime outperforms other echocardiography and medical imaging foundation models, often by a significant margin, and matches or exceeds the performance of task-specific models.

To contextualize the performance of EchoPrime and reproducibility of assessments, we assessed cardiologist interobserver variability in a series of paired echocardiogram studies from the same patient for which cardiologists noted no significant change. We found that the agreement of EchoPrime with cardiologists is comparable with the agreement between two cardiologists as EchoPrime had an average balanced accuracy of 0.89, whereas the cardiologists had a balanced accuracy of 0.82 (Extended Data Table 5). In addition, EchoPrime achieves consistently high negative predictive values across all validation sites on core tasks (Supplementary Table 6), suggesting its utility in preliminary assessment of echocardiography images as well as triaging normal studies that require more urgent clinician review.

Multi-view data improve performance

Our model synthesizes the full range of information from an echocardiogram study, including multiple videos of different views and clinical report text up to 512 tokens long. This marks a significant improvement over previous medical foundation models such as BioMedCLIP and EchoCLIP, which process only single-view, individual images and handle text up to 77 tokens long. We experimented with batch size, video encoder architecture, video weighting approaches and the number of candidate reports (Extended Data Table 6, Supplementary Tables 7 and 8 and Supplementary Fig. 2) to select the final EchoPrime hyperparameters and performance. On videos-to-text retrieval (Table 3), the correct text report appeared in the top 10 EchoPrime-matched text reports for 98% of the studies from the test set (Recall@10), outperforming EchoCLIP by 45%. Similarly, on the text-to-videos task EchoPrime achieved Recall@10 of 97%, outperforming EchoCLIP by 35%. Retrieval was more challenging in normal studies than abnormal studies, although strong performance was achieved in both subcohorts (Supplementary Table 9).

Cardiologists integrate information from multiple views and videos to provide a holistic assessment, and EchoPrime was designed to weight the interpretation of multiple videos from multiple views. To demonstrate the impact of integrating multiple clips, multiple videos and multiple views, we compared the performance of the EchoPrime encoder when input is a single frame versus single video versus multiple videos versus multiple videos with multiple instance attention weighting (Extended Data Fig. 1). The gradual improvement across multiple tasks reflects the relevance of temporal and view-specific information to each task and the importance of the anatomical attention module. EchoPrime can additionally be leveraged even in low-quality images or settings where limited views are acquired (Extended Data Table 4). We evaluated EchoPrime in the set of studies in which the study was described as 'technically difficult' in the clinician report and

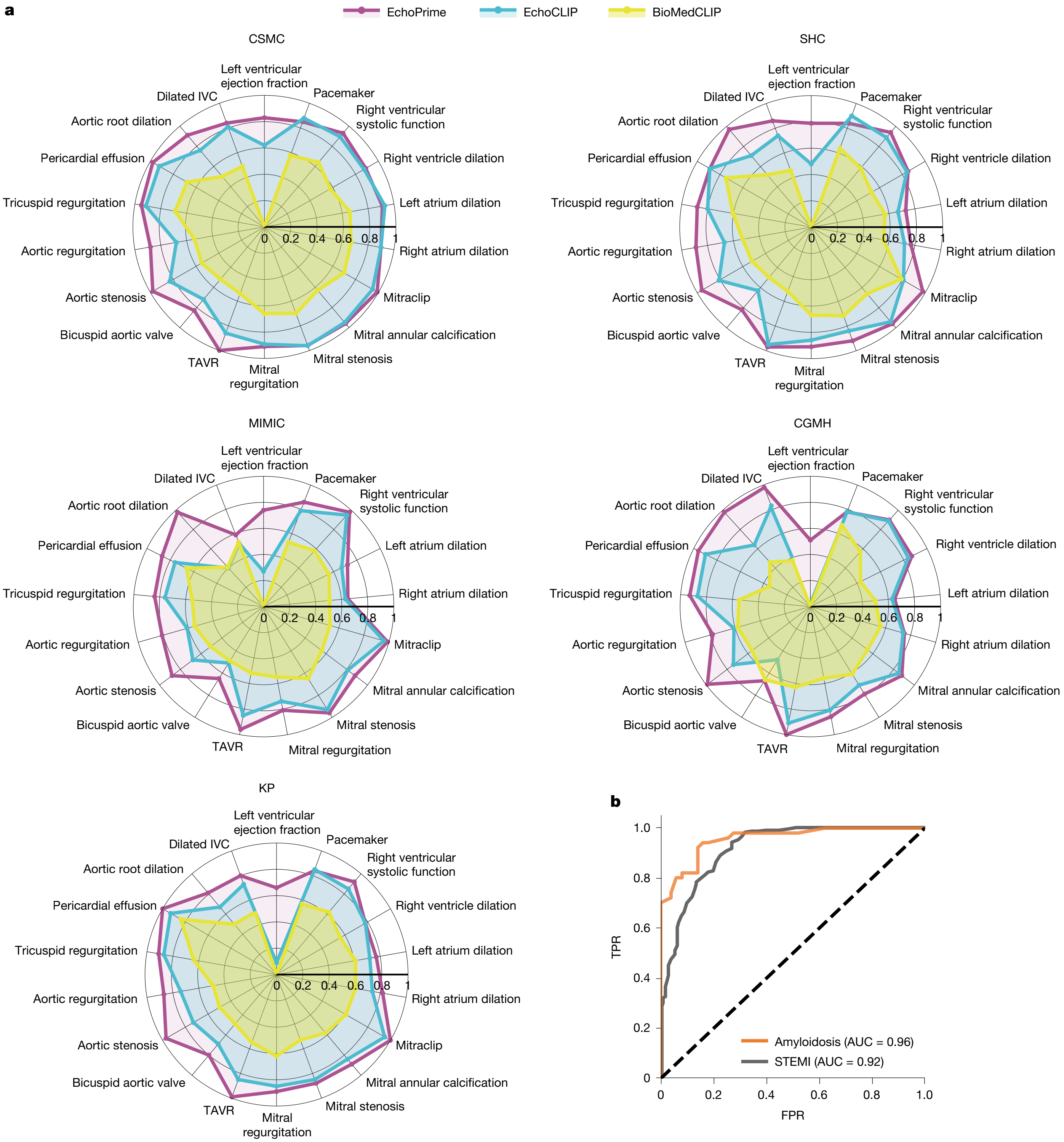


Fig. 2 | Performance metrics for echocardiography prediction tasks. **a**, Comparison of EchoPrime, EchoCLIP and BioMedCLIP on a range of echocardiography interpretation tasks on cohorts from five different clinical centres. The area under the receiver operating characteristic curve is shown for classification tasks and the R^2 score for regression tasks. **b**, Transfer learning. Using linear probing, EchoPrime accurately predicts multimodal cardiovascular diseases even when their labels are not explicitly mentioned in the echocardiography reports. FPR, false positive rate; IVC, inferior vena cava; TAVR, transcatheter aortic valve replacement; TPR, true positive rate.

show similar performance. To test the performance in a single-view setting, we tested EchoPrime on 2,172 studies with only one A4C video per study. Because other echocardiography models do not have a view classifier or use anatomical attention, a random single video per study was utilized for comparison. The performance across all tasks was similar to EchoPrime performance in the full setting.

Interpretable view weighting

Our view classifier achieves a one-versus-rest AUC of 0.997 for predicting 58 different standard echocardiographic views (Extended Data Fig. 2) on the internal test set consisting of 4,000 videos. The view prediction, as well as the underlying videos, was then used to train an anatomical

Table 3 | Cross-modal retrieval metrics on a test set of video–text report pairs

Model	Videos to text					Text to videos				
	Mean rank	Median rank	Recall@1 (%)	Recall@5 (%)	Recall@10 (%)	Mean rank	Median rank	Recall@1 (%)	Recall@5 (%)	Recall@10 (%)
EchoPrime	**2.86**	**1**	**70**	**94**	**98**	**3.05**	**1**	**69**	**94**	**97**
VideoMAE	4.24	1	67	91	95	4.44	1	66	91	95
Resnet2+1D	5.41	1	63	91	95	5.50	1	62	90	95
EchoCLIP	58.71	9	20	42	53	37.07	5	25	50	62

Lower rank is better, and higher recall is better. Best performance was for EchoPrime. EchoPrime was compared with different video encoders (VideoMAE and ResNet2+1D) trained on the same dataset.

attention module using a multiple instance attention framework to identify the most relevant views and videos for a given anatomical structure and report section (Fig. 3 and Extended Data Fig. 4). With the anatomical attention module, we enabled EchoPrime to focus on the most informative views for the anatomy being evaluated and weigh conflicting assessments of different videos from the same study. (Extended Data Fig. 2b).

For instance, in assessing the mitral valve, the model learned to focus on apical-2-chamber (A2C), A4C, parasternal short axis on the mitral valve and corresponding colour Doppler views as the key videos. Similarly, the model identified that views where the IVC is clearly visible to be the most relevant for prediction tasks involving the IVC. To provide context for these results, we asked three independent cardiologists to manually assign importance to each view, as well as identify whether each structure would be present in each view and compared the focus of the model with the consensus cardiologist focus in side-by-side weight matrices (Fig. 3d). Both EchoPrime and cardiologists show similar emphasis on important views for different tasks, such as both highlighted the suprasternal notch view for assessing the aorta and the apical views for assessing the left ventricle.

Transfer learning

We evaluated the performance of the video encoder of EchoPrime for both echocardiographic and out-of-domain medical diagnosis tasks to assess whether EchoPrime learns intrinsic knowledge about echocardiography and cardiac pathophysiology. We fine-tuned vision transformers in a low-data, fully supervised setting, comparing different initialization weights and found that initializing with EchoPrime weights leads to better performance than Kinetics initialization or training from scratch for assessment of mitral regurgitation, identification of pacemaker, diagnosis of cardiac amyloidosis and diagnosis of ST-elevation myocardial infarction (STEMI; Extended Data Fig. 3).

Multimodal cardiovascular diagnosis

Using the video embeddings in EchoPrime, linear probing can identify cardiac diseases not typically diagnosed by echocardiography with high accuracy including cardiac amyloidosis and STEMI (multimodal cardiovascular diagnoses). Using EchoPrime embeddings with linear probing, we could identify STEMI with an AUC of 0.90 and amyloidosis with an AUC of 0.95 even with limited training data. It was also possible to predict multimodal cardiovascular diagnoses without any supervised fine-tuning, with a non-parametric *k*-nearest neighbour probing approach, that could identify STEMI with an AUC of 0.92 and amyloidosis with an AUC of 0.96 (Fig. 2d).

Discussion

EchoPrime is a multi-view, view-informed, video-based deep learning algorithm for comprehensive assessment of echocardiograms. Trained on over ten times more data than existing echocardiography AI models, EchoPrime integrates view-dependent information into clinical assessments across a wide range of interpretations of cardiac structure and function. Incorporating videos from more than 10 years' worth of echocardiography data at a large academic medical centre, EchoPrime integrates anatomical attention and retrieval augmented interpretation to achieve state-of-the-art performance on a wide range of interpretation tasks beyond both previous foundation models and task-specific echocardiography AI models. In addition, EchoPrime excels in many tasks where no previous AI models have been developed and shows generalizability in four external validation cohorts. With publicly released code, weights and demo, we hope EchoPrime can be a resource to the medical and AI communities.

Compared with previous echocardiographic models, EchoPrime has a longer temporal context, longer text context length and integrates complementary videos with a robust view classifier and multiple instance anatomical attention. Trained on more than ten times the data of other recent echocardiography AI models[24,31], EchoPrime improves upon models in low-resource and low-quality image settings. By utilizing multiple instance learning to integrate information from multiple videos, we reproduced the relationship between echocardiographic views and the associated anatomical structures that cardiologists use to make EchoPrime an inherently clinically interpretable model. Trained on the largest corpus of echocardiographic data, our model can detect rare diseases and cardiovascular diseases not typically assessed by cardiac ultrasound[31,32].

A few limitations are worth considering. As EchoPrime relies on a video encoder, still images such as spectral Doppler and M-mode were excluded from the training set. Even though much of their information is already present in B-mode and colour Doppler videos, future iterations might benefit from including these image types. Parallel work in direct annotation of echocardiography images can help to inform assessment of echocardiographic measurements[31] (Supplementary Table 10). Although EchoPrime was evaluated at four geographically distinct academic medical centres, all data were collected retrospectively. Future experiments should include prospective evaluations in real-world settings to understand human–computer interaction as well as accuracy and acceptability of EchoPrime assessment. Working within a complex ecosystem such as health care requires understanding the human–computer interaction through clinical trials and identifying optimal integration into clinical workflows. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information. Combined with AI-based probe guidance[33,34], EchoPrime has the potential to automate the echocardiographic workflow.

The Achilles heel of medical imaging lies in human heterogeneity, and opportunities in the future lie in the potential integration of AI into the health-care system. Our results represent an important step towards the automated evaluation of cardiac ultrasound. EchoPrime augments existing methods for interpreting echocardiography and has the potential to streamline manual tasks and improve the reproducibility of cardiac imaging. Trained on the most echocardiography data to date and using anatomical attention as well as retrieval augmented interpretation, EchoPrime performs comprehensive evaluation of echocardiography videos.

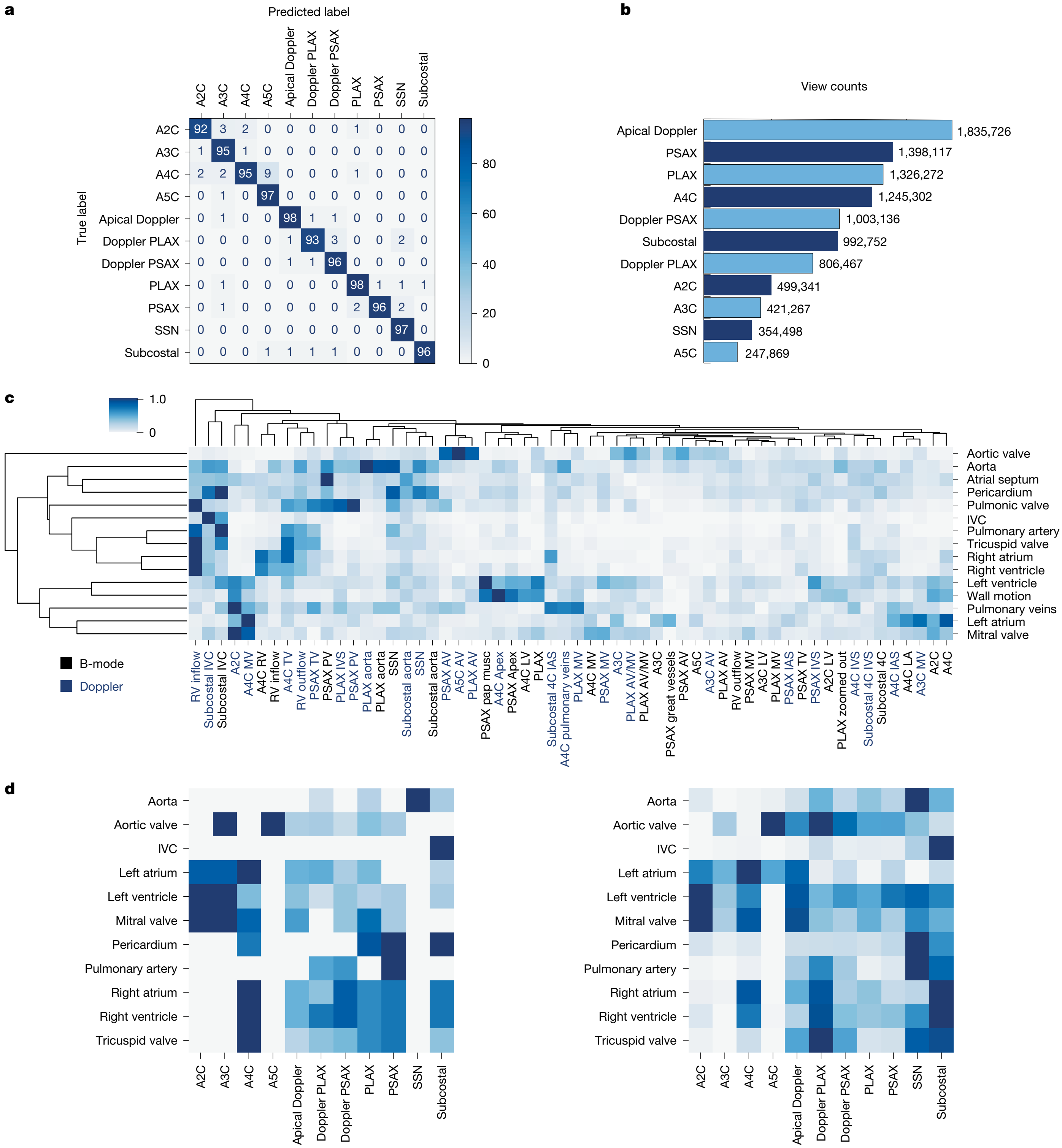


Fig. 3 | Anatomical attention to weight predictions across videos in an echocardiogram study. a, A view classifier was trained on 60,000 videos to distinguish between 58 standard echocardiographic views. **b**, A comprehensive echocardiogram examination comprises many different views, which is summarized across our training cohort. **c**, A clustered heatmap showing the relative priority and ranking of each video for each anatomical structure based on learned anatomical attention. IAS, interatrial septum; IVS, interventricular septum; LA, left atrium; LV, left ventricle; MV, mitral valve; PV, pulmonic valve; RV, right ventricle; TV, tricuspid valve. **d**, Comparison of how a cardiologist (left) and EchoPrime (right) prioritizes different views based on the assessed anatomical structure.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-09850-x.

1. ACCF/ASE/AHA/ASNC/HFSA/HRS/SCAI/SCCM/SCCT/SCMR. 2011 Appropriate use criteria for echocardiography. *J. Am. Soc. Echocardiogr.* **24**, 229–267 (2011).

2. Zhang, J. et al. Fully automated echocardiogram interpretation in clinical practice. *Circulation* **138**, 1623–1635 (2018).
3. Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Vision–language foundation model for echocardiogram interpretation. *Nat. Med.* **30**, 1481–1488 (2024).
4. Zhang, S. et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI* **2**, AIoa2400640 (2025).
5. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
6. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
7. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
8. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th Intl Conf. Machine Learning* (eds Meila, M. & Zhang, T.) 8748–8763 (PMLR, 2021).
9. Jia, C. et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. 38th Intl Conf. Machine Learning* (eds Meila, M. & Zhang, T.) 4904–4916 (PMLR, 2021).
10. Li, J., Li, D., Xiong, C. & Hoi, S. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. 39th Intl Conf. Machine Learning* (eds Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G. & Sabato, S.) 12888–12900 (PMLR, 2022).
11. Singh, A. et al. FLAVA: a foundational language and vision alignment model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 15617–15629 (2022).
12. Li, H. et al. Uni-Perceiver v2: a generalist model for large-scale vision and vision-language tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2691–2700 (IEEE, 2023).
13. Alayrac, J.-B. et al. Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **35**, 23716–23736 (2022).
14. Wang, W. et al. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 19175–19186 (IEEE, 2023).
15. Zhai, X., Mustafa, B., Kolesnikov, A. & Beyer, L. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 11941–11952 (IEEE, 2023).
16. Lavoie, S. et al. Modeling caption diversity in contrastive vision-language pretraining. In *Proc. 41st Intl Conf. Machine Learning* (eds Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J. & Berkenkamp, F.) 26070–26084 (PMLR, 2024).
17. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
18. Huang, K. et al. A foundation model for clinician-centered drug repurposing. *Nat. Med.* https://doi.org/10.1038/s41591-024-03233-x (2024).
19. Bannur, S. et al. Learning to exploit temporal structure for biomedical vision-language processing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 15016–15027 (IEEE, 2023).
20. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* https://doi.org/10.1038/s41586-023-06555-x (2023).
21. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
22. Eslami, S., Meinel, C. & de Melo, G. PubMedCLIP: how much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023* (eds Vlachos, A. & Augenstein, I.) 1181–1193 (Association for Computational Linguistics, 2023).
23. Lin, W. et al. PMC-CLIP: contrastive language-image pre-training using biomedical documents. In *Proc. MICCAI 2023* (eds Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T. & Taylor, R.) 525–536 (Springer, 2023).
24. Holste, G. et al. Complete AI-enabled echocardiography interpretation with multitask deep learning. *JAMA* https://doi.org/10.1001/jama.2025.8731 (2025).
25. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
26. Vrudhula, A. et al. Automated deep learning phenotyping of tricuspid regurgitation in echocardiography. *JAMA Cardiol*. **10**, 595–602 (2025).
27. Vrudhula, A. et al. High-throughput deep learning detection of mitral regurgitation. *Circulation* **150**, 923–933 (2024).
28. Yıldız Potter, İ., Leo, M. M., Vaziri, A. & Feldman, J. A. Automated detection and localization of pericardial effusion from point-of-care cardiac ultrasound examination. *Med. Biol. Eng. Comput.* **61**, 1947–1959 (2023).
29. Holste, G. et al. Severe aortic stenosis detection by deep learning applied to echocardiography. *Eur. Heart J.* **44**, 4592–4604 (2023).
30. Ghorbani, A. et al. Deep learning interpretation of echocardiograms. *Npj Digit. Med.* **3**, 10 (2020).
31. Sahashi, Y. et al. Artificial intelligence automation of echocardiographic measurements. *JACC* **86**, 964–978 (2025).
32. Tromp, J. et al. A formal validation of a deep learning-based automated workflow for the interpretation of the echocardiogram. *Nat. Commun.* **13**, 6776 (2022).
33. Shida, Y., Kumagai, S., Tsumura, R. & Iwata, H. Automated image acquisition of parasternal long-axis view with robotic echocardiography. *IEEE Robot. Autom. Lett.* **8**, 5228–5235 (2023).
34. Jiang, H. et al. Cardiac Copilot: automatic probe guidance for echocardiography with world model. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* (eds Linguraru, M. G. et al.) 190–199 (Springer Nature, 2024).

Methods

Dataset and cohort curation

The CSMC echocardiography laboratory performs clinical echocardiography for a wide range of indications ranging from asymptomatic pre-operative screening to evaluation for open heart surgery or heart transplant. A standard full resting echocardiogram study consists of a series of 50–150 videos and images visualizing the heart from different angles and locations, using multiple acquisition techniques including standard 2D videos, tissue Doppler images and colour Doppler videos. Each echocardiogram study corresponds to a unique patient during a specific examination, with different videos representing different facets of cardiac function. We utilized all adult transthoracic echocardiogram studies for EchoPrime training, excluding studies for paediatric echocardiography and adult congenital heart disease echocardiograms.

For EchoPrime training, we curated a dataset of 12,124,168 2D and colour Doppler echocardiography videos from 275,442 studies and corresponding medical reports from 108,913 patients collected between 2011 and 2022. DICOM images were queried from the clinical data storage system, converted to AVI video files, and de-identified before model training and inference. Data were split by patient into training, validation and internal test datasets. The training data contained 11,984,170 videos derived from 272,256 studies across 107,663 patients from the CSMC, and the CSMC data were the only data used to train the model. The validation data contained 25,167 videos from 565 studies across 250 patients and were used to select the best-model checkpoint. The internal test set contained 114,831 videos derived from 2,621 studies across 1,000 patients, it was completely withheld during training and used solely for a single, final evaluation of our model. The full training dataset consists of transthoracic echocardiogram (TTEs) studies, transoesophageal echocardiogram and stress echocardiography studies; however, downstream experiments and validation tests focused specifically on TTEs, which accounted for 234,036 out of a total of 275,442 studies.

For linear probing and transfer learning experiments, we compiled two focused datasets for diseases not typically diagnosed by echocardiography (multimodal cardiovascular diagnoses): one for STEMI and another for amyloidosis. Studies from the CSMC were collected and reviewed, with clinical expert labelling of 641 STEMI cases and 159 cardiac amyloidosis cases. For the STEMI dataset, all reviewed studies included ECG modality, and positive cases were determined through conventional clinical definition (ST changes on ECG). For cardiac amyloidosis, all patients had confirmatory technetium pyrophosphate, cardiac MRI, biopsy or laboratory testing. To ensure representation in the test set, approximately one-third of the cases were reserved for testing (200 STEMI and 50 amyloidosis), whereas the remaining cases were used for training. To assess inter-physician variability in echocardiography interpretations, we evaluated clinician interpretations and EchoPrime interpretations on a cohort of patients undergoing consecutive echocardiogram studies for which a cardiologist thought there was no significant difference between the two studies. This dataset comprised 4,607 paired studies for which we evaluated the degree of variation between the first and second studies of the same patient.

External validation datasets

For external validation, we obtained echocardiographic data from four geographically distinct international sites to evaluate the performance of EchoPrime. From SHC, we collected 91,746 videos from 1,792 TTE studies across 1,792 patients. From the MIMIC dataset, we used 75,768 videos from 2,431 TTE studies with corresponding echocardiography reports across 1,767 patients. From the CGMH Hospital in Taiwan, we acquired 24,724 videos with reports from 1,188 TTE studies, each from a unique patient. The CGMH dataset had 21 videos per study on average, notably, smaller from the three US sites (31–51). From the KP medical centre, we collected 201,752 videos from 4,891 studies obtain in September 2023. All video data from external datasets underwent the same preprocessing workflow, including RGB conversion, cropping to the ultrasound region and resizing to 224 × 224. Study labels were extracted in quantitative form from reports using regex with institution-specific custom phrases. All data from external validation sites were used exclusively for validation and were not accessible to the model during training.

Contrastive vision–language training

To develop video and text encoders that map videos and text into a joint representation space (Fig. 1b), we adopted the contrastive strategy used in CLIP[8], but modified the encoder architectures to allow for video and long context text input. For the video encoder, we selected a multi-scale vision transformer (mVIT)[35] architecture with weights pretrained on the kinetics dataset[36]. The last layer of the video encoder was modified such that the output was 512-dimensional. For the text encoder, we selected Wordpiece Encoding tokenization and BERT[37] architecture with weights pretrained on PubMed abstracts (BioMedBERT[38]). A fully connected layer was appended on top of the BERT classification (CLS) token vector representation to obtain a similarly sized 512-dimensional embedding. mVIT was chosen as the video encoder because mVIT effectively captures temporal dynamics, as evidenced by performance changes when video frames are shuffled, a phenomenon not observed in other vision transformers[39]. PubMedBERT was selected as the text encoder due to its pretraining on 21 Gb of medical text, giving our model a head start in understanding medical literature. PubMedBERT has a maximum context length of 512 tokens (which is of sufficient length to capture 99.56% of the reports in our training dataset).

During training, batches of 32 video–report pairs were constructed, with each pair consisting of a video and report from the same study, and all pairs within a batch coming from different studies. The input video size was 224 × 224 × 16 × 3 with a stride of 2 to increase the temporal context despite only 16 frames. The video data pixel intensity was normalized using the mean and standard deviation of the training dataset. For reports longer than 512 tokens, a random 512 tokens were selected starting with either a start token or the first token appearing after a [SEP] token. Videos were fed through the video encoder and reports were fed through the text encoder to obtain embeddings. A dot product was taken between video and report embeddings to obtain a similarity matrix. This similarity matrix was scaled by a trainable temperature parameter (starting at 1.0).

The cross-entropy loss was calculated between this scaled similarity matrix and an identity matrix, which represents the correct matching of video–reports pairs. The AdamW optimizer is used, with a starting learning rate of 4×10^{-5}, weight decay of 1×10^{-6} and the ReduceOnPlateau learning rate scheduler. We froze the weights of the first six layers of the text encoder to retain the general knowledge captured by the pretrained BioMedBERT, while updating all other encoder parameters. The best checkpoint was selected based on the validation loss. With this setup, the model was first pretrained on the full dataset for 60 epochs. It was then fine-tuned on a refined dataset, which included cleaned interpretation text reports that corrected common typographical errors and excluded stress echocardiogram and transoesophageal echocardiogram studies for 20 epochs. Multiple batch sizes were compared during training with similar performance (Supplementary Table 7).

Learning anatomical view weighting

Different echocardiography views offer different insights into cardiac form and function, with some cardiac structures being only visible in certain views but all views providing complementary diagnostic information. We utilized an attention-based deep multiple instance learning (MIL)[40] method to learn relative weights that captures the importance of each echocardiogram view and video for interpretations of different anatomical structures.

We first trained a view classifier to provide a fine-grained characterization of echocardiogram views as defined by the American Society of Echocardiography guidelines[41]. We selected 77,426 echocardiogram videos, a subset of videos from the training set, for cardiac sonographers to label into 58 different view categories and trained an image-based ConvNextBase[42] view classification model on 224 × 224 images. The AdamW optimizer was used to minimize the cross-entropy loss, and the best checkpoint was selected based on validation loss. During training random image frames were taken from each labelled video and augmentation techniques including RandAugment[43] and RandomErasing[44] were applied. Held-out datasets of studies from different patients were used for validation and testing.

For MIL, we used the contrastive video encoder and view classifier to produce a concatenated embedding for each input video of a given echocardiogram study (Extended Data Fig. 4). The concatenated embedding is a one-hot-encoded view classification vector as well as the contrastive video embedding, and MIL is trained to produce importance weights for each video in relation to interpretations separated by anatomical structure. These MIL-generated weights were then used to compute a weighted average of the video embeddings produced by the video encoder. The resulting weighted average and individual video-level prediction were passed through a multilayer perception (MLP) to produce an ensembled final prediction. During training, the prediction was compared with the ground truth to compute loss, which was then backpropagated to update the MLP and MIL weights. The output vector of alphas for MIL represents the importance of each view for assessing each anatomical structure. The anatomical structure was defined by structured section in the echocardiogram report ('left ventricle', 'right ventricle', and so on) and each interpretation task was ensembled to create total weighting vectors for each anatomical section, forming the anatomical attention module.

Cross-modal retrieval

We assessed the retrieval accuracy of EchoPrime at the study level. For each echocardiogram study, we generated embeddings for all available videos using the video encoder and averaged to create a study embedding. Similarly, using the text encoder, the report text was mapped to a report text embedding. Using these EchoPrime embeddings, we performed a cross-modal search to find text reports or studies that are semantically similar to a given query study or a query text report by their cosine similarity. To evaluate accuracy, we reported the mean and median rank of the correct candidate for both videos-to-report and report-to-videos retrieval. In addition, we followed the approach used in ALIGN[9] and measured Recall@K, which indicates the percentage of the test set for which the correct result is found within the top *K*-retrieved samples. For comparison, we also calculated retrieval metrics for the previously developed EchoCLIP model.

Retrieval augmented interpretation

Similar to retrieval augmented generation[21], we leveraged the corpus of historical echocardiogram reports and videos for EchoPrime study interpretation. Using pre-trained parametric memory (EchoPrime weights) and non-parametric memory (a corpus of clinical reports), our approach leveraged anatomical attention to weight the cross-modal retrieval based on each section in an echocardiogram report. Interpretations were weighted by anatomy and video to obtain comprehensive echocardiography examination interpretations. The anatomical attention module identifies the most informative views for the specified section and computes a weighted average of video embeddings based on their views, resulting in a section-specific study embedding. We used this method to generate 50 candidate interpretation reports and average features across these 50 reports to obtain final predictions. Quantitative assessments are the average of the 50 reports, whereas the final EchoPrime report displayed the finding if an individual finding was present in more than half of the retrieved reports for categorial findings. Hyperparameter $k = 50$ was chosen because it achieves the optimal trade-off between performance and runtime (Supplementary Fig. 2). This approach does not require additional fine-tuning and can generalize to predict any feature present in the reports.

Multimodal disease diagnosis

We applied *k*-nearest neighbour and linear probing on top of EchoPrime for multimodal cardiovascular diagnoses prediction. We focused on STEMI and cardiac amyloidosis, for which echocardiography is not the definitive diagnostic test but can provide corroborating evidence. In this setting, patients with clinically diagnosed STEMI or cardiac amyloidosis were identified and the EchoPrime embeddings were used to evaluate whether linear probing can identify patients with this diagnosis despite not being in the echo report text. Study embeddings are obtained by averaging the embeddings of all videos within each study and used along with clinician-assigned diagnoses to fit the models. For *k*-nearest neighbour probing, we used KNeighborsClassifier, and for linear probing, we applied LogisticRegression, both from scikit-learn library.

Impact of different initialization

We tested whether initializing the model with EchoPrime weights leads to higher label efficiency and accuracy than initializing with kinetics weights and random initialization. We selected four binary classification tasks: pacemaker detection, mitral regurgitation classification, amyloidosis detection and STEMI detection. Pacemaker detection was trained on A4C right ventricle views, mitral regurgitation on Doppler A4C mitral valve views, amyloid detection on PLAX views and STEMI detection was trained on A4C views. All tasks were trained on datasets of 32, 64, 128 and 256 samples, with an equal distribution of positive and negative samples.

Evaluation and benchmarking

To evaluate the accuracy of echocardiography interpretation, we chose 19 cardiac features of different anatomical structures, including 17 binary and 2 continuous features (Supplementary Table 11). We identified corresponding report phrases in report text to obtain ground-truth labels based on the presence or absence of a specific interpretation or numerical measurements. For tasks assessing severity (trivial, mild, moderate and severe), we binarized assessments by considering a feature as present if the severity was classified as moderate or severe. Accuracy for classification tasks was measured with the area under the receiver operating characteristic curve and coefficient of determination (R^2 score) for regression tasks. To calculate the confidence intervals for performance metrics, we used a bootstrapping technique. Specifically, we randomly sampled the dataset with replacement, maintaining the same sample size as the original dataset, and computed the performance metrics. This process was repeated 10,000 times to generate a distribution of the metrics. Finally, we sorted these bootstrapped metrics and determined the 2.5th and 97.5th percentiles to obtain the 95% confidence intervals. To assess the accuracy of our multi-label view classifier, we used the average area under the receiver operating characteristic curve using a one-vs-rest approach, treating the correct class as the positive label and all other classes as the negative label and averaged the results across all classes.

To compare the performance with other foundation models, we also obtained interpretations using EchoCLIP and BioMedCLIP using retrieval. We selected one A4C video per study, embedded each frame of the video and averaged the embedding to produce a study embedding. Then, similarly to EchoPrime retrieval, we found the closest reports to this study embedding and averaged their features to make final predictions. We used this method of retrieval because although both EchoCLIP and BioMedCLIP are image-based models, A4C is considered the most informative view by physicians (Fig. 3d), and this approach follows the methodology outlined in the EchoCLIP paper[3].

Computing hardware and software

We used Python (v3.8.13), PyTorch (v2.1.2, CUDA 12.1; https://pytorch.org) and TorchVision (v0.17.0) for all experiments and analyses in the study. We used scikit-learn (v1.2.0) (https://scikit-learn.org/) for probing methods, umap-learn (v0.5; https://umap-learn.readthedocs.io) for dimensionality reduction and scipy (v1.12.0; https://scipy.org/) for statistics and linear algebra operations. To train EchoPrime, we used two 50-Gb NVIDIA RTX A6000 GPUs configured for multi-GPU training using PyTorch's DistributedDataParallel. We obtained weights of previously developed foundation models from the Hugging Face model hub (https://huggingface.co/docs/hub/en/models-the-hub): EchoCLIP (https://huggingface.co/mkaichristensen/echo-clip-r) and BioMedCLIP (https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224).

Ethics statement

This study used retrospective echocardiographic data obtained under data-use agreements with CSMC, SHC, KP and CGMH. Data were de-identified before analysis. Use of the MIMIC-IV-Echo dataset was conducted under its data-use agreement and approved access through PhysioNet. No prospective data collection or human participant intervention was performed.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Echocardiographic data from the MIMIC, used as an external validation site, are available as the MIMIC-IV-Echo dataset (https://physionet.org/content/mimic-iv-echo/0.1/), where credentialed researchers can apply for access. Data from other participating sites (CSMC, SCH, CGMH and KP) are not publicly available and can be accessed through data-use agreements with the respective institutions.

Code availability

The code used for model inference and the trained model weights are publicly available on GitHub (https://github.com/echonet/EchoPrime).

35. Fan, H. et al. Multiscale vision transformers. In *Proc. IEEE/CVF Intl Conf. Computer Vision (CVPR)* 6824–6835 (2021).
36. Carreira, J. & Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (IEEE, Honolulu, HI, 2017).
37. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (eds Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, 2019).
38. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2022).
39. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. In *Intl Conf. Learning Representations* (ICLR, 2021).
40. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *Proc. 35th Intl Conf. Machine Learning* (eds Dy, J. & Krause, A.) 2127–2136 (PMLR, 2018).
41. Mitchell, C. et al. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the American Society of Echocardiography. *J. Am. Soc. Echocardiogr.* **32**, 1–64 (2019).
42. Liu, Z. et al. A ConvNet for the 2020s. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11976–11986 (2022).
43. Cubuk, E. D., Zoph, B., Shlens, J. & Le, Q. RandAugment: practical automated data augmentation with a reduced search space. *Adv. Neural Inf. Process. Syst.* **33**, 18613–18624 (2020).
44. Zhong, Z., Zheng, L., Kang, G., Li, S. & Yang, Y. Random erasing data augmentation. *Proc. AAAI Conf. Artif. Intell.* **34**, 13001–13008 (2020).

Acknowledgements D.O. discloses US National Institutes of Health NHLBI grants R00HL157421, R01HL173526 and R01HL173487.

Author contributions M.V., B.H. and D.O. conceived the study, developed and trained the deep learning models, and performed the statistical analyses and result interpretation. M.V., I.-M.C., X.T. and T.Y.C. prepared and curated the echocardiography datasets. D.O., N.Y., P.C., D.L. and S.C. contributed to the clinical interpretation of results. M.V., B.H. and D.O. wrote the manuscript with input from all authors.

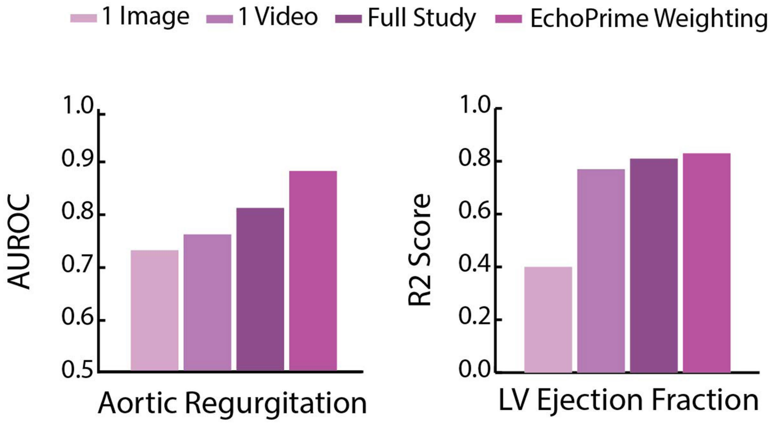
Competing interests The authors declare no competing interests.

Additional information

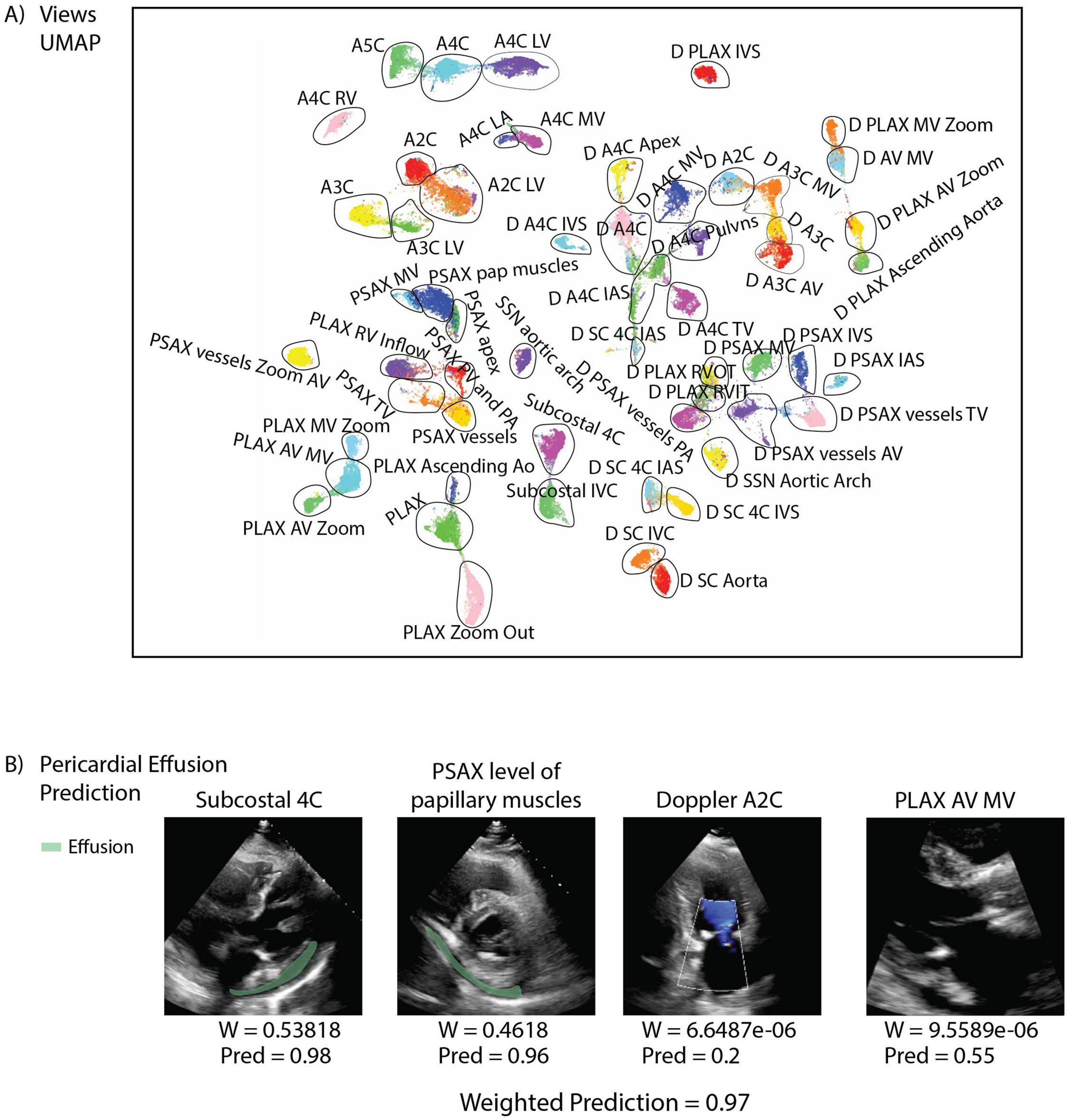
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-09850-x.

Correspondence and requests for materials should be addressed to Bryan He or David Ouyang.

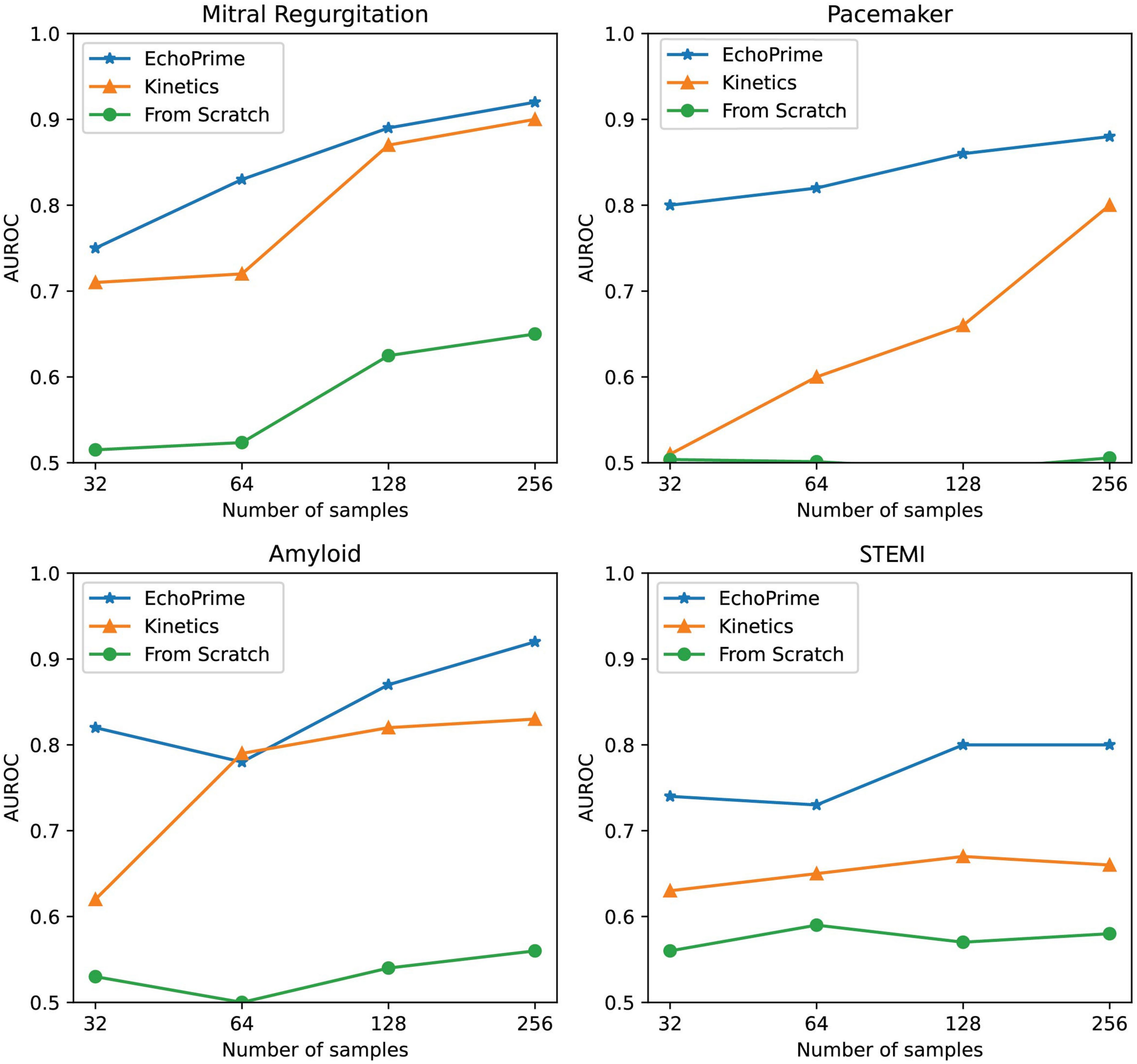
Peer review information *Nature* thanks Tyler Smith, James Thomas and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



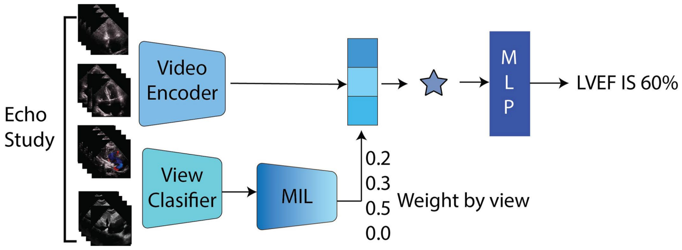
Extended Data Fig. 1 | The accuracy of predicting aortic regurgitation severity and left ventricular ejection fraction improves progressively from predictions on a single image to a full video, a single video to multiple videos in a full study and finally applying anatomical attention (EchoPrime weighting).



Extended Data Fig. 2 | A, UMAP representation of features from the second to last layer of the video classifier. Doppler views are denoted by prefix "D". **B**, Case study on pericardial effusion prediction demonstrates the advantages of utilizing all videos from the study and applying anatomical weighting.



Extended Data Fig. 3 | Comparison of model performance when initialized with EchoPrime weights versus Kinetics weights versus random weights, showing the impact of domain-specific initialization on training efficiency and convergence.



Extended Data Fig. 4 | Weighting the views based on relevance. After training the architecture shown in the figure, with view classifier and video encoder frozen, MIL learns which views are most informative views for the specific feature prediction. The MIL component can be isolated and used as a weighting vector in other applications.

Extended Data Table 1 | EchoPrime versus task-specific models

	ECHOPRIME INTERNAL	TASK SPECIFIC INTERNAL	ECHOPRIME EXTERNAL	TASK SPECIFIC EXTERNAL
LV ejection fraction (R^2)	**0.83 (0.81-0.85)**	0.81	**0.79 (0.76-0.82)**	0.77
LV ejection fraction (MAE)	4.79 (4.60-4.97)	**4.10**	**4.14 (3.94-4.34)**	6.00
Pacemaker	0.85 (0.82-0.88)	-	0.84 (0.81-0.87)	-
RV systolic function	0.93 (0.91-0.95)	-	0.94 (0.88-1.00)	-
RV dilation	0.89 (0.83-0.95)	-	0.85 (0.80-0.90)	-
LA dilation	**0.91 (0.86-0.96)**	0.85	0.73 (0.70-0.75)	-
RA dilation	0.90 (0.82-0.97)	-	0.77 (0.74-0.81)	-
Mitraclip	0.99 (0.99-1.00)	-	0.98 (0.94-1.00)	-
Mitral annular calcification	0.96 (0.95-0.97)	-	0.96 (0.94-0.98)	-
Mitral stenosis	0.96 (0.91-0.99)	-	0.92 (0.86-0.98)	-
Mitral regurgitation	**0.92 (0.91-0.94)**	0.92 (0.90-0.93)	0.91 (0.89-0.93)	**0.95 (0.92–0.97)**
TAVR	1.00 (0.99-1.00)	-	0.97 (0.90-1.00)	-
Bicuspid AV	0.83 (0.67-0.96)	-	0.82 (0.73-0.90)	-
Aortic stenosis	**0.98 (0.96-0.99)**	**0.98 (0.97- 0.99)**	**0.96 (0.92-0.99)**	0.95 (0.94- 0.96)
Aortic regurgitation	0.88 (0.83-0.93)	-	0.89 (0.80-0.97)	-
Tricuspid regurgitation	**0.95 (0.93-0.97)**	0.93 (0.91- 0.94)	0.88 (0.86-0.91)	**0.95 (0.94 - 0.96)**
Pericardial effusion	**0.98 (0.95-0.99)**	0.94 (0.75- 0.99)	0.89 (0.77-0.98)	-
Aortic root dilation	0.91 (0.83-0.98)	-	0.97 (0.94-0.99)	-
Dilated IVC	0.84 (0.82-0.86)	-	0.86 (0.81-0.91)	-
PA pressure (R^2)	0.43 (0.39-0.47)	-	0.36 (0.30-0.42)	-
PA pressure (MAE)	7.30 (6.95-7.67)	-	7.97 (7.35-8.66)	-

Extended Data Table 2 | EchoPrime's performance on regression tasks across three medical centres

Task	CSMC	KP	SHC
Ejection Fraction R^2	0.83 (0.82-0.85)	0.66 (0.62-0.71)	0.79 (0.76-0.82)
Ejection Fraction MAE (%)	4.64 (4.47-4.81)	4.51 (4.30-4.72)	4.14 (3.94-4.34)
Pulmonary Artery Pressure R^2	0.42 (0.36-0.46)	0.31 (0.27-0.35)	0.36 (0.30-0.42)
Pulmonary Artery Pressure MAE (mmHg)	7.30 (6.96-7.65)	7.39 (7.08-7.70)	7.97 (7.35-8.66)
Wall Motion Score R^2	0.77 (0.74-0.80)	-	-
Wall Motion Score MAE (unitless)	0.10 (0.10-0.11)	-	-
TAPSE R^2	0.39 (0.26-0.51)	0.28 (0.02-0.46)	0.34 (0.00–0.70)
TAPSE MAE (cm)	0.31 (0.27-0.35)	0.35 (0.30-0.40)	0.25 (0.13-0.33)
MV Area R^2	0.34 (0.19-0.44)	-	0.22 (0.16-0.28)
MV Area MAE (cm^2)	0.78 (0.69-0.88)	-	1.03 (0.62-1.44)
Transaortic velocity R^2	0.70 (0.59-0.78)	0.50 (0.34-0.61)	0.54 (0.23-0.75)
Transaortic velocity MAE (cm/s)	20.4 (18.1-22.8)	56.4 (49.7-63.7)	51.1 (30.6-75.2)
Transaortic gradient R^2	0.69 (0.64-0.74)	0.50 (0.45-0.55)	0.38 (0.33-0.53)
Transaortic gradient MAE (mmHg)	5.00 (4.62-5.40)	6.02 (5.56 – 6.51)	7.32 (6.63-8.00)
TR gradient R^2	0.36 (0.30-0.41)	0.26 (0.12-0.43)	0.28 (0.23-0.32)
TR gradient MAE (mmHg)	6.72 (6.39-7.07)	6.33 (5.41-7.26)	7.41 (6.93-7.85)
Sinus of Valsalva Size R^2	0.62 (0.55-0.67)	0.39 (0.30-0.47)	-
Sinus of Valsalva Size MAE (cm)	0.24 (0.22-0.26)	0.33 (0.31-0.35)	-
Ascending Aorta Size R^2	0.36 (0.23-0.48)	0.34 (0.23-0.44)	0.26 (0.20-0.32)
Ascending Aorta Size MAE (cm)	0.31 (0.28-0.34)	0.34 (0.32-0.36)	0.40 (0.37-0.43)
IVC Diameter R^2	0.44 (0.38-0.49)	-	0.18 (0.14-0.23)
IVC Diameter MAE (cm)	0.29 (0.28-0.30)	-	0.41 (0.39-0.43)

Extended Data Table 3 | Aggregated comparison of EchoPrime and PanEcho on all external validation sites. Asterisks (*) denote statistically significant differences (p < 0.05/59 adjusted for multiplicity). The DeLong test was used for comparing AUCs, and bootstrapping was used for MAE and R2 metrics

	STANFORD		**MIMIC**		**CGMH**		**KP**	
	EchoPrime	PanEcho	EchoPrime	PanEcho	EchoPrime	PanEcho	EchoPrime	PanEcho
LV ejection fraction (R^2)	**0.79 (0.76-0.82)***	0.39 (0.36-0.42)	**0.74 (0.66-0.80)***	0.25 (0.20-0.30)	**0.51 (0.44-0.57)***	0.12 (0.04-0.19)	**0.66 (0.62-0.71)***	0.28 (0.23-0.33)
LV ejection fraction (MAE)	**4.14 (3.94-4.34)***	6.41 (6.03-6.79)	**6.21 (5.54-6.92)***	10.95 (9.91-12.01)	**6.46 (6.19-6.74)***	9.02 (8.69-9.35)	**4.51 (4.30-4.72)***	6.66 (6.37-6.97)
Pacemaker	0.84 (0.81-0.87)	-	**0.86 (0.80-0.92)**	-	**0.78 (0.63-0.92)**	-	0.84 (0.81-0.87)	---
RV systolic function	**0.94 (0.88-1.00)**	0.93 (0.80-0.99)	**0.98 (0.97-0.99)***	0.85 (0.78-0.93)	0.90 (0.79-0.98)	**0.93 (0.86-0.98)**	**0.92 (0.88-0.95)***	0.86 (0.82-0.90)
RV dilation	0.85 (0.80-0.90)	**0.87 (0.83-0.90)**	-	-	**0.87 (0.69-0.97)***	0.69 (0.50-0.87)	**0.78 (0.73-0.83)**	0.77 (0.72-0.82)
LA dilation	0.73 (0.70-0.75)	**0.81 (0.79-0.83)***	**0.72 (0.68-0.75)***	0.68 (0.65-0.71)	0.65 (0.62-0.67)	**0.72 (0.69-0.76)***	**0.77 (0.75-0.79)***	0.75 (0.73-0.77)
RA dilation	0.77 (0.74-0.81)	**0.84 (0.81-0.87)***	0.65 (0.61-0.69)	**0.70 (0.66-0.73)***	0.75 (0.67-0.84)	**0.87 (0.81-0.92)***	0.82 (0.79-0.84)	**0.84 (0.82-0.87)**
Mitraclip	**0.98 (0.94-1.00)**	-	**0.99 (0.99-1.00)**	-	-	-	**1.00 (1.00-1.00)**	-
Mitral annular calcification	**0.96 (0.94-0.98)**	-	**0.88 (0.85-0.91)**	-	**0.88 (0.75-0.96)**	-	**0.89 (0.86-0.91)**	-
Mitral stenosis	0.92 (0.86-0.98)	**0.96 (0.94-0.98)**	**0.96 (0.95-0.98)**	0.85 (0.66-0.99)	0.79 (0.59-0.98)	**0.88 (0.79-0.95)**	0.88 (0.78-0.95)	**0.92 (0.87-0.96)**
Mitral regurgitation	**0.91 (0.89-0.93)**	0.85 (0.79-0.90)	**0.81 (0.79-0.83)***	0.73 (0.71-0.76)	**0.86 (0.82-0.89)***	0.76 (0.70-0.82)	**0.89 (0.86-0.91)***	0.76 (0.72-0.80)
TAVR	**0.97 (0.90-1.00)**	---	**0.96 (0.94-0.98)**	-	**1.00 (1.00-1.00)**	-	**0.99 (0.98-0.99)**	---
Bicuspid AV	**0.82 (0.73-0.90)***	0.73 (0.64-0.82)	0.65 (0.52-0.78)	**0.66 (0.50-0.81)**	**0.67 (0.45-1.00)**	0.47 (0.20-0.80)	**0.80 (0.72-0.88)***	0.72 (0.63-0.79)
Aortic stenosis	0.96 (0.92-0.99)	**0.97 (0.95-0.98)**	**0.88 (0.84-0.91)***	0.80 (0.77-0.84)	**0.99 (0.98-1.00)**	0.92 (0.86-0.97)	**0.97 (0.96-0.98)***	0.87 (0.85-0.90)
Aortic regurgitation	**0.89 (0.80-0.97)**	0.83 (0.74-0.90)	**0.81 (0.75-0.86)***	0.69 (0.63-0.76)	**0.78 (0.70-0.84)**	0.77 (0.70-0.83)	**0.87 (0.83-0.92)***	0.73 (0.68-0.78)
Tricuspid regurgitation	**0.88 (0.86-0.91)**	0.85 (0.81-0.88)	**0.84 (0.82-0.86)***	0.76 (0.73-0.78)	**0.93 (0.90-0.95)***	0.88 (0.84-0.91)	**0.91 (0.89-0.93)***	0.87 (0.84-0.90)
Pericardial effusion	**0.89 (0.77-0.98)**	0.87 (0.73-0.96)	**0.87 (0.83-0.92)***	0.80 (0.75-0.86)	0.96 (0.91-1.00)	**0.99 (0.98-1.00)**	**1.00 (0.99-1.00)**	0.98 (0.94-1.00)
Aortic root dilation	**0.97 (0.94-0.99)**	0.80 (0.73-0.85)	**0.98 (0.96-0.99)**	0.60 (0.29-0.91)	**0.98 (0.97-1.00)***	0.72 (0.58-0.84)	**0.81 (0.78-0.85)***	0.58 (0.54-0.63)
Dilated IVC	**0.86 (0.81-0.91)**	-	**0.59 (0.49-0.69)**	-	**0.98 (0.94-1.00)**	-	**0.80 (0.77-0.82)**	-
PA pressure (R^2)	**0.36 (0.30-0.42)***	0.07 (0.00-0.14)	-	-	-	-	**0.31 (0.27-0.35)***	0.16 (0.13-0.19)
PA pressure (MAE)	**7.97 (7.35-8.66)***	9.12 (8.32-9.96)	-	-	-	-	**7.39 (7.08-7.70)***	8.07 (7.73-8.41)

Extended Data Table 4 | EchoPrime performance in low-quality resource-constrained setting

	FULL TEST		LOW-QUALITY		SINGLE-VIEW	
	EchoPrime	PanEcho	EchoPrime	PanEcho	EchoPrime	PanEcho
LV ejection fraction (R^2)	**0.83 (0.81-0.85)**	0.52 (0.50-0.54)	**0.80 (0.75-0.84)**	0.42 (0.37-0.46)	**0.78 (0.75-0.80)**	0.34 (0.29-0.38)
LV ejection fraction (MAE)	**4.79 (4.60-4.97)**	7.46 (7.16-7.78)	**5.30 (4.86-5.76)**	8.98 (8.22-9.77)	**5.22 (5.02-5.43)**	8.49 (8.12-8.87)
Pacemaker	**0.85 (0.82-0.88)**	---	**0.89 (0.84-0.93)**	---	**0.83 (0.80-0.85)**	---
RV systolic function	**0.93 (0.91-0.95)**	0.89 (0.87-0.92)	**0.91 (0.85-0.96)**	0.85 (0.77-0.92)	**0.92 (0.89-0.94)**	0.83 (0.79-0.86)
RV dilation	**0.89 (0.83-0.95)**	0.86 (0.79-0.91)	0.85 (0.72-0.96)	**0.88 (0.78-0.95)**	**0.92 (0.87-0.95)**	0.76 (0.68-0.83)
LA dilation	**0.91 (0.86-0.96)**	0.89 (0.86-0.92)	**0.95 (0.89-0.99)**	0.88 (0.77-0.97)	**0.92 (0.87-0.96)**	0.80 (0.75-0.85)
RA dilation	0.90 (0.82-0.97)	**0.93 (0.89-0.96)**	**0.89 (0.70-1.00)**	0.85 (0.75-0.95)	**0.90 (0.82-0.97)**	0.88 (0.80-0.93)
Mitraclip	**0.99 (0.99-1.00)**	---	**1.00 (0.99-1.00)**	---	**0.97 (0.94-0.99)**	---
Mitral annular calcification	**0.96 (0.95-0.97)**	---	**0.97 (0.94-0.98)**	---	**0.94 (0.91-0.96)**	---
Mitral stenosis	**0.96 (0.91-0.99)**	**0.96 (0.93-0.98)**	0.97 (0.90-1.00)	**0.99 (0.97-1.00)**	**0.94 (0.87-0.98)**	0.79 (0.69-0.87)
Mitral regurgitation	**0.92 (0.91-0.94)**	0.86 (0.84-0.89)	**0.91 (0.84-0.96)**	0.85 (0.75-0.93)	**0.89 (0.87-0.91)**	0.77 (0.74-0.81)
TAVR	**1.00 (0.99-1.00)**	---	**1.00 (0.99-1.00)**	---	**0.90 (0.87-0.93)**	---
Bicuspid AV	0.83 (0.67-0.96)	**0.95 (0.89-0.98)**	0.72 (0.44-1.00)	**0.84 (0.66-1.00)**	**0.78 (0.62-0.92)**	0.76 (0.65-0.86)
Aortic stenosis	**0.98 (0.96-0.99)**	0.93 (0.90-0.95)	**0.93 (0.83-1.00)**	0.92 (0.81-0.98)	0.80 (0.75-0.85)	**0.86 (0.82-0.90)**
Aortic regurgitation	**0.88 (0.83-0.93)**	0.77 (0.71-0.83)	**0.78 (0.38-0.99)**	0.73 (0.44-0.97)	**0.75 (0.68-0.82)**	0.73 (0.66-0.80)
Tricuspid regurgitation	**0.95 (0.93-0.97)**	0.91 (0.88-0.94)	**0.98 (0.97-0.99)**	0.95 (0.90-0.99)	**0.92 (0.88-0.94)**	0.84 (0.80-0.88)
Pericardial effusion	**0.98 (0.95-0.99)**	0.93 (0.89-0.96)	**0.94 (0.91-0.97)**	0.92 (0.83-0.98)	**0.93 (0.89-0.97)**	0.80 (0.73-0.86)
Aortic root dilation	**0.91 (0.83-0.98)**	0.66 (0.56-0.76)	**0.96 (0.91-1.00)**	0.51 (0.23-0.74)	**0.81 (0.70-0.91)**	0.67 (0.54-0.79)
Dilated IVC	**0.84 (0.82-0.86)**	---	**0.82 (0.77-0.87)**	---	**0.80 (0.78-0.83)**	---
PA pressure (R^2)	**0.43 (0.39-0.47)**	0.27 (0.23-0.30)	**0.32 (0.17-0.44)**	0.16 (0.05-0.25)	**0.36 (0.31-0.40)**	0.03 (-0.04-0.10)
PA pressure (MAE)	**7.30 (6.95-7.67)**	8.27 (7.88-8.68)	**7.76 (6.96-8.63)**	8.91 (8.05-9.81)	**7.67 (7.29-8.06)**	9.55 (9.10-10.03)

Extended Data Table 5 | EchoPrime versus physician variability compared to physician versus physician

N=4607	ECHOPRIME VS PHYSICIAN	PHYSICIAN VS PHYSICIAN
LV ejection fraction (R^2)	**0.83 (0.81-0.85)**	0.73 (0.70-0.75)
LV ejection fraction (MAE)	4.79 (4.60-4.97)	**4.59 (4.47-4.72)**
Pacemaker	0.78 (0.75-0.80)	**0.83 (0.80-0.85)**
RV systolic function	**0.88 (0.85-0.90)**	0.82 (0.79-0.86)
RV dilation	**0.87 (0.81-0.92)**	0.71 (0.66-0.76)
LA dilation	**0.86 (0.81-0.92**)	0.77 (0.72-0.82)
RA dilation	**0.90 (0.84-0.96)**	0.76 (0.69-0.84)
Mitraclip	**0.98 (0.95-0.99)**	0.96 (0.93-0.98)
Mitral annular calcification	**0.90 (0.87-0.92)**	0.78 (0.75-0.81)
Mitral stenosis	**0.92 (0.91-0.93)**	0.70 (0.59-0.82)
Mitral regurgitation	0.86 (0.83-0.88)	**0.87 (0.85-0.90)**
TAVR	0.98 (0.98-0.99)	**0.99 (0.98-0.99)**
Bicuspid AV	0.84 (0.70-0.96)	**0.87 (0.81-0.92)**
Aortic stenosis	**0.95 (0.93-0.96)**	0.83 (0.78-0.88)
Aortic regurgitation	**0.83 (0.78-0.88)**	0.75 (0.69-0.80)
Tricuspid regurgitation	**0.90 (0.88-0.93)**	0.86 (0.83-0.89)
Pericardial effusion	**0.91 (0.87-0.94)**	0.88 (0.81-0.94)
Aortic root dilation	**0.87 (0.79-0.94)**	0.77 (0.71-0.82)
PA pressure (R^2)	0.43 (0.39-0.47)	**0.50 (0.44-0.55)**
PA pressure (MAE)	7.30 (6.95-7.67)	**5.89 (5.70-6.09)**
Wall Motion Score (R^2)	**0.77 (0.74-0.80)**	0.76 (0.72-0.80)
Wall Motion Score (MAE)	0.10 (0.10-0.11)	**0.04 (0.04-0.05)**
MV Area (R^2)	0.34 (0.19-0.44)	**0.38 (0.13-0.59)**
MV Area (MAE)	0.78 (0.69-0.88)	**0.70 (0.60-0.81)**
Transaortic gradient (R^2)	0.69 (0.64-0.74)	**0.82 (0.78-0.86)**
Transaortic gradient (MAE)	5.00 (4.62-5.40)	**3.12 (2.97-3.27)**
TRgradient (R^2)	0.36 (0.30-0.41)	**0.50 (0.33-0.67)**
TRgradient (MAE)	6.72 (6.39-7.07)	**5.42 (5.21-5.66)**
Sinus of Valsalva Size (R^2)	**0.62 (0.55-0.67)**	0.42 (0.06-0.87)
Sinus of Valsalva Size (MAE)	0.24 (0.22-0.26)	**0.21 (0.16-0.30)**
Ascending Aorta Size (R^2)	0.36 (0.23-0.48)	**0.81 (0.77-0.85)**
Ascending Aorta Size (MAE)	0.31 (0.28-0.34)	**0.17 (0.16-0.19)**
IVC Diameter (R^2)	**0.44 (0.38-0.49)**	0.05 (-0.68-0.28)
IVC Diameter (MAE)	**0.29 (0.28-0.30)**	0.32 (0.31-0.34)

Extended Data Table 6 | EchoPrime performance with and without anatomical weighting variability

	CSMC		SHC	
	Without	With	Without	With
LV ejection fraction (R^2)	0.80 (0.78-0.82)	**0.83 (0.81-0.85)**	0.78 (0.75-0.81)	**0.79 (0.76-0.82)**
LV ejection fraction (MAE)	4.97 (4.78-5.17)	**4.79 (4.60-4.97)**	4.22 (4.01-4.42)	**4.14 (3.94-4.34)**
Pacemaker	0.82 (0.80-0.85)	**0.85 (0.82-0.88)**	0.80 (0.77-0.83)	**0.84 (0.81-0.87)**
RV systolic function	0.92 (0.90-0.94)	**0.93 (0.91-0.95)**	0.91 (0.80-1.00)	**0.94 (0.88-1.00)**
RV dilation	**0.90 (0.85-0.95)**	0.89 (0.83-0.95)	0.83 (0.77-0.88)	**0.85 (0.80-0.90)**
LA dilation	0.91 (0.86-0.96)	0.91 (0.86-0.96)	0.71 (0.68-0.73)	**0.73 (0.70-0.75)**
RA dilation	**0.92 (0.85-0.98)**	0.90 (0.82-0.97)	0.76 (0.73-0.80)	**0.77 (0.74-0.81)**
Mitraclip	0.98 (0.96-0.99)	**0.99 (0.99-1.00)**	0.97 (0.93-1.00)	**0.98 (0.94-1.00)**
Mitral annular calcification	0.95 (0.93-0.96)	**0.96 (0.95-0.97)**	0.96 (0.93-0.98)	**0.96 (0.94-0.98)**
Mitral stenosis	0.96 (0.90-0.99)	0.96 (0.91-0.99)	**0.93 (0.86-0.98)**	0.92 (0.86-0.98)
Mitral regurgitation	0.91 (0.89-0.93)	**0.92 (0.91-0.94)**	0.89 (0.85-0.93)	**0.91 (0.89-0.93)**
TAVR	0.99 (0.99-1.00)	**1.00 (0.99-1.00)**	**0.99 (0.98-1.00)**	0.97 (0.90-1.00)
Bicuspid AV	0.82 (0.67-0.96)	**0.83 (0.67-0.96)**	0.78 (0.69-0.87)	**0.82 (0.73-0.90)**
Aortic stenosis	0.96 (0.93-0.98)	**0.98 (0.96-0.99)**	0.94 (0.89-0.99)	**0.96 (0.92-0.99)**
Aortic regurgitation	0.83 (0.77-0.88)	**0.88 (0.83-0.93)**	0.84 (0.72-0.94)	**0.89 (0.80-0.97)**
Tricuspid regurgitation	0.94 (0.91-0.96)	**0.95 (0.93-0.97)**	0.84 (0.79-0.88)	**0.88 (0.86-0.91)**
Pericardial effusion	0.95 (0.92-0.98)	**0.98 (0.95-0.99)**	0.88 (0.75-0.98)	**0.89 (0.77-0.98)**
Aortic root dilation	0.88 (0.79-0.96)	**0.91 (0.83-0.98)**	0.94 (0.90-0.98)	**0.97 (0.94-0.99)**
Dilated IVC	0.82 (0.80-0.84)	**0.84 (0.82-0.86)**	0.82 (0.74-0.88)	**0.86 (0.81-0.91)**
PA pressure (R^2)	0.41 (0.37-0.45)	**0.43 (0.39-0.47)**	0.33 (0.27-0.39)	**0.36 (0.30-0.42)**
PA pressure (MAE)	7.44 (7.10-7.82)	**7.30 (6.95-7.67)**	**7.89 (7.21-8.60)**	7.97 (7.35-8.66)

Reported values are balanced accuracies.

nature portfolio

Corresponding author(s): David Ouyang

Last updated by author(s): 10/17/2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected from the echocardiography laboratories at Cedars-Sinai Medical Center, Stanford Healthcare, Chang Gung Memorial Hospital and Kaiser-Permanente. Additionally, publicly available echocardiography dataset, MIMIC-IV-ECHO, from Beth Israel Deaconess Medical Center was downloaded. Only data from Cedars-Sinai was used for training the machine learning model, all other sites were reserved for external validation. Medical images were converted from initial DICOM files into AVI video files prior to deep learning training. Associated text and patient identifiers were mapped from the electronic healthcare record for training but videos were de-identified prior to input into AI model.
Data analysis	A deep learning algorithm was used to assess the echocardiogram videos. Retrieval augmented interpretation was performed to obtain comprehensive echocardiography interpretations. Our code and working model is available at https://github.com/echonet/EchoPrime. All analysis code was written in Python v3.8.13. and required software packages for model training and inference include PyTorch (v2.1.2, CUDA 12.1), TorchVision (v0.17.0), Scikit Learn (v1.2.0), Umap Learn(v0.5) and Scipy (1.12.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Echocardiography data from the Beth Israel Deaconess Medical Center, used as an external validation site, is available as the MIMIC-IV-Echo dataset at https://physionet.org/content/mimic-iv-echo/0.1/, where credentialed researchers can apply for access. Data from other participating sites (Cedars-Sinai, Stanford, CGMH, and Kaiser Permanente) is not publicly available due to its potentially identifiable nature and can be accessed through data use agreements with the respective institutions.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender

Cohort demographics including patient sex are described in Table 1.

Reporting on race, ethnicity, or other socially relevant groupings

The input data for training comes from a large academic medical center with diverse patient population (demographics shown in Table 1). The race, ethnicity and other socially relevant groupings were not used for model input given recognized biases that might happen if that were an input predictor (Duffy et al. npj Digital Medicine)

Population characteristics

Echocardiograms acquired at Cedars Sinai Medical Center between 2011 and 2022 were used to train the model. Detailed test and training cohort information in Table 1.

Recruitment

A waiver of consent was obtained for the use of retrospective de-identified data. Patient data from 2011 to 2022 were used in de-identified format without prospective recruitment.

Ethics oversight

This research was approved by the Cedars-Sinai Medical Center (Study00001409), Kaiser Permanente Northern California (2238961) and Stanford Healthcare Institutional Review Boards (Study 43721). A waiver of consent was obtained for the use of retrospective de-identified data.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

12,124,168 echocardiogram videos were collected from 275,442 unique studies performed on 108,913 patients. Sample size calculations were not done prior to the development of the model as we sought to optimize for the largest possible training dataset size. The sample size was chosen based on the availability of echocardiogram videos for training at the healthcare site. Given that prior echocardiogram AI models are trained on much less data (10-100x smaller training dataset sizes), we anticipate use of greater than 10 million samples would be sufficient to train a foundation model.

For external validation, we used echocardiography datasets from multiple institutions, including Stanford Healthcare (91,746 videos from 1,792 patients), Beth Israel Deaconess Medical Center via the MIMIC-IV-Echo dataset (75,768 videos from 2,431 patients), Chang Gung Memorial Hospital in Taiwan (24,724 videos from 1,188 patients), and Kaiser Permanente Medical Center (201,752 videos from 4,891 patients). No formal sample size calculation was performed for the external datasets, as we aimed to include the maximal amount of data available from each institution. The dataset sizes were considered sufficient to evaluate model performance on key clinical metrics such as AUROC and R2, given that each site contained an adequate number of positive and negative cases for all core outcomes assessed.

Data exclusions

No data was excluded from model training.

Replication

95% confidence intervals were calculated using bootstrapping. The algorithm otherwise is deterministic and code is available. Model evaluation was independently performed across four external validation datasets, and results were consistently robust across sites.

Randomization

Patients were randomly divided into training, validation and testing splits with approximate ratio of 95:1:4.

Blinding

This study was a retrospective analysis of existing echocardiography data and corresponding reports obtained from clinical archives. As no new data were collected and all analyses were performed using de-identified datasets and automated algorithms, there were no experimental conditions or human assessors to blind. Blinding was therefore not applicable to this study design.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern
☒	□ Plants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.