nature medicine

Article

https://doi.org/10.1038/s41591-024-02959-y

Vision–language foundation model for echocardiogram interpretation

Received: 17 August 2023

Accepted: 28 March 2024

Published online: 30 April 2024

Check for updates

Matthew Christensen[1], Milos Vukadinovic[1,2], Neal Yuan[3,4] & David Ouyang[1,5]

The development of robust artificial intelligence models for echocardiography has been limited by the availability of annotated clinical data. Here, to address this challenge and improve the performance of cardiac imaging models, we developed EchoCLIP, a vision–language foundation model for echocardiography, that learns the relationship between cardiac ultrasound images and the interpretations of expert cardiologists across a wide range of patients and indications for imaging. After training on 1,032,975 cardiac ultrasound videos and corresponding expert text, EchoCLIP performs well on a diverse range of benchmarks for cardiac image interpretation, despite not having been explicitly trained for individual interpretation tasks. EchoCLIP can assess cardiac function (mean absolute error of 7.1% when predicting left ventricular ejection fraction in an external validation dataset) and identify implanted intracardiac devices (area under the curve (AUC) of 0.84, 0.92 and 0.97 for pacemakers, percutaneous mitral valve repair and artificial aortic valves, respectively). We also developed a long-context variant (EchoCLIP-R) using a custom tokenizer based on common echocardiography concepts. EchoCLIP-R accurately identified unique patients across multiple videos (AUC of 0.86), identified clinical transitions such as heart transplants (AUC of 0.79) and cardiac surgery (AUC 0.77) and enabled robust image-to-text search (mean cross-modal retrieval rank in the top 1% of candidate text reports). These capabilities represent a substantial step toward understanding and applying foundation models in cardiovascular imaging for preliminary interpretation of echocardiographic findings.

Echocardiography, or cardiac ultrasound, is the most common, non-invasive method of evaluating heart function and identifying heart disease. Echocardiography routinely guides clinical cardiology decision-making[1–3] and is used for disease diagnosis, risk stratification and assessment of treatment response[1,4]. Recent work has used artificial intelligence (AI) to improve the accuracy of echocardiographic measurements[5–7] and disease diagnoses[8–10]; however, these AI approaches focus on narrow individual tasks that require specific training for each task and do not use vision–language foundation models[11].

Recent advances in AI have leveraged representation learning on large image and text datasets to develop vision–language foundation models that generalize beyond narrow sets of predefined tasks[12,13].

[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 Medicine, University of California San Francisco, San Francisco, CA, USA. [4]Division of Cardiology, San Francisco Veterans Affairs Medical Center, San Francisco, CA, USA. [5]Division of Artificial Intelligence in Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA. e-mail: david.ouyang@cshs.org

https://doi.org/10.1038/s41591-024-02959-y

Table 1 | Clinical characteristics of Cedars-Sinai Medical Center study cohort, reported per echocardiography study

	Total	Training	Validation	Test
n	224,685	195,082	8,119	21,484
Age (mean (s.d.))	66.26 (16.74)	66.3 (16.7)	65.8 (17.0)	65.7 (16.9)
Female=true (%)	96,451 (42.9)	83,700 (42.9)	3,363 (41.4)	9,388 (43.7)
Race (%)				
Native American	526 (0.2)	456 (0.2)	23 (0.3)	47 (0.2)
Asian	16,601 (7.5)	14,450 (7.5)	555 (6.9)	1,596 (7.5)
Black	29,546 (13.3)	25,624 (13.3)	1,104 (13.8)	2,818 (13.3)
Hispanic	22,424 (10.1)	19,394 (10.0)	842 (10.5)	2,188 (10.3)
Non-Hispanic white	133,399 (60.0)	116,044 (60.1)	4,699 (58.6)	12,656 (59.6)
Other	15,376 (6.9)	13,243 (6.9)	612 (7.6)	1,521 (7.2)
Pacific Islander	767 (0.3)	688 (0.4)	36 (0.4)	43 (0.2)
Unknown	3,700 (1.7)	3,182 (1.6)	149 (1.9)	369 (1.7)
AF	46,994 (20.9)	41,214 (21.1)	1,633 (20.1)	4,147 (19.3)
HF	75,358 (33.5)	65,802 (33.7)	2,764 (34.0)	6,792 (31.6)
HTN	90,738 (40.4)	79,229 (40.6)	3,250 (40.0)	8,259 (38.4)
CVA/TIA/TE	38,283 (17.0)	33,475 (17.2)	1,378 (17.0)	3,430 (16.0)
MI	14,983 (6.7)	13,120 (6.7)	514 (6.3)	1,349 (6.3)
CAD	55,659 (24.8)	48,840 (25.0)	2,040 (25.1)	4,779 (22.2)
PAD	23,369 (10.4)	20,475 (10.5)	838 (10.3)	2,056 (9.6)
DM	37,900 (16.9)	33,226 (17.0)	1,351 (16.6)	3,323 (15.5)
CKD	40,947 (18.2)	35,960 (18.4)	1,482 (18.3)	3,505 (16.3)
Previous smoker	7,632 (3.4)	6,593 (3.4)	256 (3.2)	783 (3.6)

AF, atrial fibrillation; HF, heart failure; HTN, hypertension; CVA, cerebrovascular accident; TIA, transient ischemic attack; TE, thromboembolism; MI, myocardial infarction; CAD, coronary artery disease; PAD, pulmonary artery disease; DM, diabetes mellitus; CKD, chronic kidney disease.

Table 2 | Main performance metrics

	Image encoder	Tokenizer	MCMRR	LVEF, MAE	PAP, MAE	TAVR, AUC	MitraClip, AUC	Pacemaker, AUC
CLIP	ViT-B-32	CLIP BPE	10,743.0	20.8 (20.7–20.8)	16.8 (16.8–16.9)	0.46 (0.46–0.47)	0.53 (0.52–0.54)	0.51 (0.51–0.52)
EchoCLIP	ConvNeXt	CLIP BPE	571.3	**8.4 (8.3–8.4)**	**10.8 (10.8–10.9)**	**0.92 (0.91–0.92)**	**0.97 (0.97–0.97)**	**0.84 (0.84–0.84)**
EchoCLIP-R (full-report prompts)	ConvNeXt	Template tokenizer	**206.1**	10.9 (10.9–11.0)	13.2 (13.1–13.2)	0.85 (0.85–0.86)	0.95 (0.94–0.95)	0.77 (0.77–0.78)
EchoCLIP-R (base)	ConvNeXt	Template tokenizer	**206.1**	16.9 (16.8–17.0)	17.5 (17.4–17.5)	0.52 (0.51–0.52)	0.81 (0.81–0.82)	0.66 (0.65–0.66)

Retrieval ranks are out of 21,484 candidates. Performance of the best-performing model for each metric is bolded. Ranges in parentheses indicate 95% CI bootstrapped with 1,000 random samples. MCMRR, mean cross-modal retrieval rank; BPE, Byte-Pair Encoding.

These models learn to encode images and text into compact representations that can then be used to perform a wide variety of separate prediction tasks for which the model was never specifically trained ('zero-shot' tasks). Given the broad range of data used to train these models, the performance of foundation models are often more robust than with conventional convolutional neural networks[14,15]. In biomedical applications, foundation models have been developed to organize biological[16–18] and medical[19] datasets, including modality-specific models for chest X-rays, retinal imaging, wearable waveforms and pathology images[20–25]. Training of foundation models on medical imaging has been bottlenecked by dataset size and is often limited to publicly available data that may not represent the range of disease severities and possible presentations. While text information might be imprecise, clinician evaluations of medical imaging provide an information-rich distillation of complex data.

In this work, we introduce EchoCLIP, a foundation model for echocardiography trained on a dataset of 1,032,975 echocardiogram videos sourced from over a decade of clinical imaging. We developed a method for substantially compressing echocardiography reports, simplifying the matching of clinical text assessments to images to focus on important clinical concepts. To assess the model's performance, we tested the model's ability to assess cardiac function, pulmonary artery pressure (PAP) and chamber size, as well as identify common intracardiac devices in both held-out internal test cohorts as well as external test cohorts. By using the model to compare pairs of echocardiogram studies, we can assess the model's ability to identify unique patients across time, identify clinically important changes in disease state and retrieve relevant clinical text for given images. Finally, we propose a new vision–language model interpretation approach based on matching relevant text with important regions of interest in images.

https://doi.org/10.1038/s41591-024-02959-y

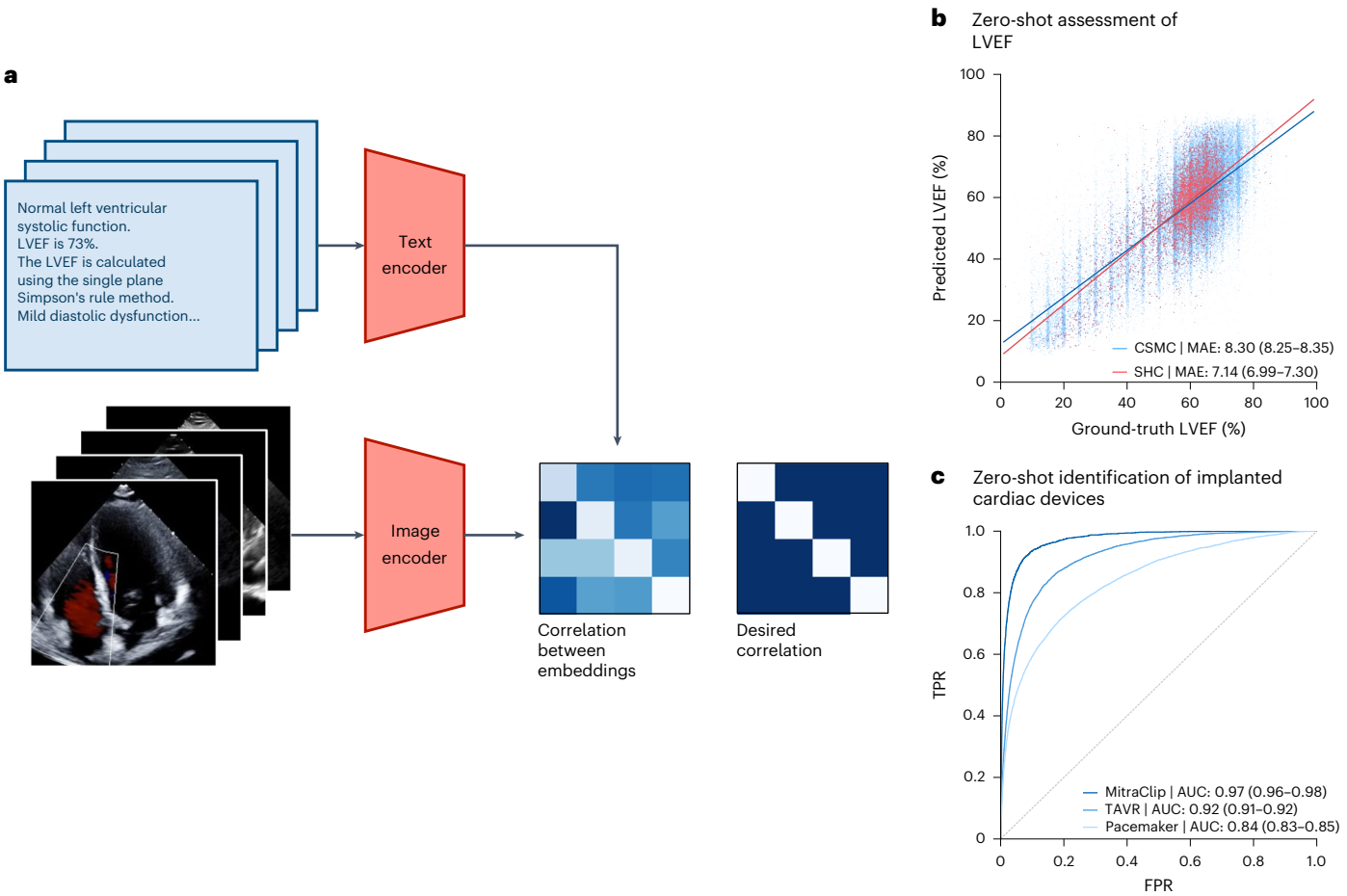


Fig. 1 | EchoCLIP workflow. a, EchoCLIP is a foundation model trained on more than 1 million echocardiogram videos across 11 years. It is composed of an image encoder for processing echocardiogram video frames and a text encoder for processing the corresponding physician interpretations. These two encoders project the images and interpretations onto a joint embedding space. **b**, Scatterplot of zero-shot prediction versus label of left ventricular ejection fraction (LVEF) in held-out test dataset from Cedars-Sinai Medical Center (CSMC; blue, $n = 100{,}994$) and Stanford Healthcare (SHC; red, $n = 5{,}000$). **c**, AUC performance for various implanted intracardiac devices, including MitraClip, TAVR valves and implanted pacemaker/defibrillator on held-out test dataset from Cedars-Sinai Medical Center. FPR, false positive rate; TPR, true positive rate.

Results

EchoCLIP is an echocardiography vision–language model trained with 1,032,975 video–text pairs derived from 224,685 echocardiography studies across 99,870 patients across a decade of clinical care (Table 1). In a self-supervised approach, EchoCLIP is trained on pairs of echocardiogram images (randomly sampled from video frames) and associated clinical report text without direct labeling of clinical interpretations or measurements. The EchoCLIP model uses a ConvNeXt-Base[26] image encoder and a Byte-Pair Encoding text tokenizer[27]. The text encoder architecture is a decoder-only transformer identical to the architecture used by the original CLIP paper[23] and has an input context length of 77 tokens. Despite not being directly trained on specific interpretation tasks, EchoCLIP can accurately identify implanted devices as well as assess cardiac form and function (Table 2). To assess the importance of pretraining and architecture[28], different architectures and dataset configurations were compared (Supplementary Table 1).

To fit an entire echocardiography report into the text encoder, a domain-specific echocardiography text tokenization format succinctly summarizing common cardiovascular concepts was developed. The model variant trained with this tokenization format, EchoCLIP-R, is capable of retrieving relevant clinical text from images and characterizes clinical changes over time. We also introduce a saliency mapping approach based on cosine similarity, PromptCAM, to show that EchoCLIP prioritizes important image features relevant to the associated text. This approach identifies clinically relevant regions of interest in echocardiography images based on prompted clinical text.

Echocardiogram interpretation without supervised learning

Without fine-tuning or task-specific training, we evaluated EchoCLIP's performance on a wide range of benchmark classification tasks in our internal held-out test set. EchoCLIP can accurately identify intracardiac devices, including percutaneous mitral valve repair with an AUC of 0.97 (95% CI 0.97–0.98), transvenous aortic valve replacement (TAVR) with an AUC of 0.92 (95% CI 0.91–0.92) and pacemaker/defibrillator leads with an AUC of 0.84 (95% CI 0.84–0.85). EchoCLIP can also detect changes from a healthy cardiac chamber size, including severe dilation of the right ventricle with an AUC of 0.92 (95% CI 0.91–0.92), right atrium with an AUC of 0.97 (95% CI 0.97–0.98), left ventricle with an AUC of 0.92 (95% CI 0.92–0.93) and left atrium with an AUC of 0.91 (95% CI 0.90–0.92). Last, EchoCLIP can assess for tamponade (AUC 0.96, 95% CI 0.94–0.98) and severe left ventricular hypertrophy (AUC 0.82, 95% CI 0.81–0.83). The sensitivity and specificity for each task are described in Extended Data Table 1. Performance was similar across key subsets stratified by age, sex and image quality (Supplementary Table 2).

External validation of cardiac function and pressure assessment

We further evaluated EchoCLIP's performance on quantitative tasks, including evaluation of left ventricular ejection fraction (LVEF) and

https://doi.org/10.1038/s41591-024-02959-y

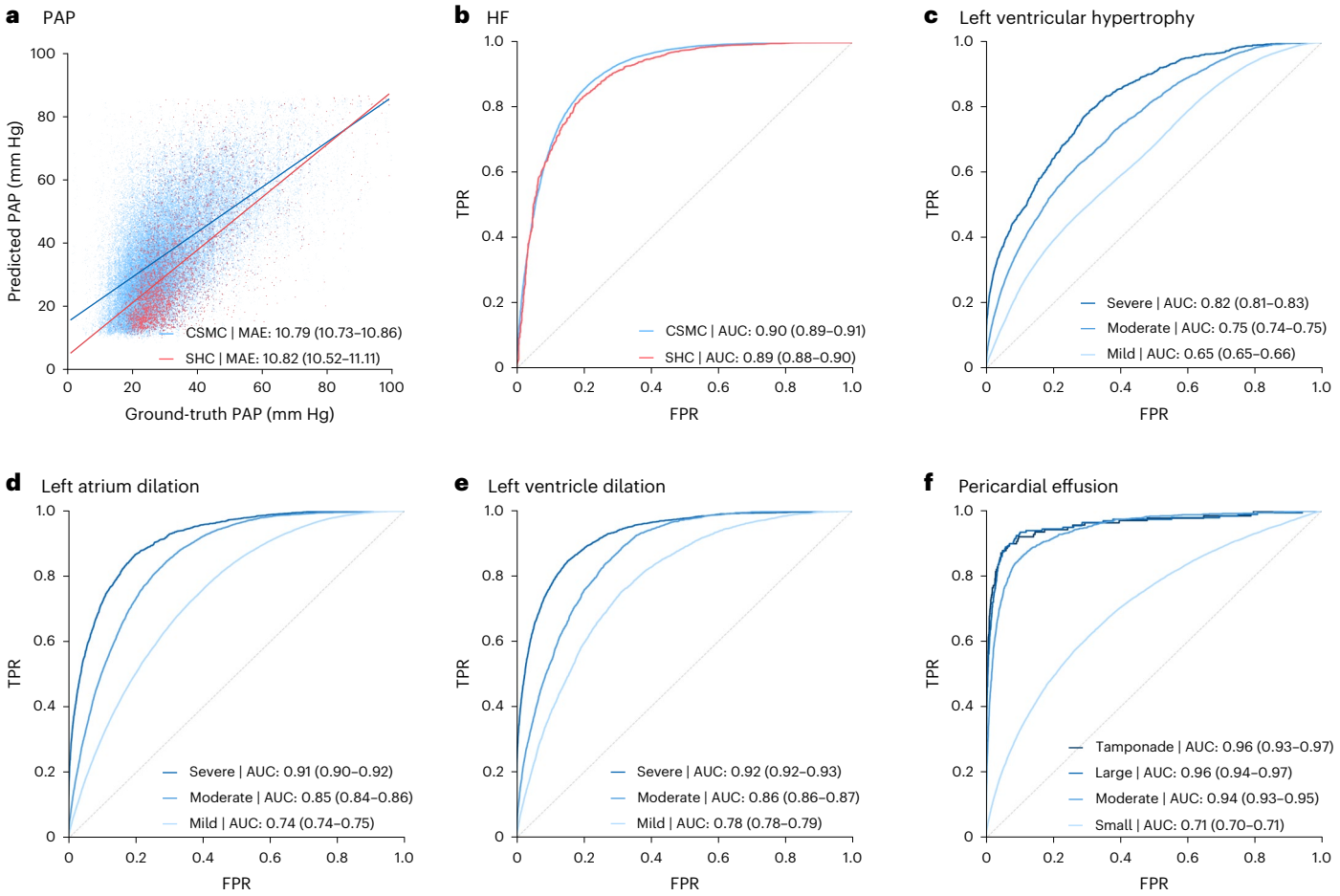


Fig. 2 | Zero-shot model performance on held-out test apical-four-chamber videos. a, Estimation of pulmonary artery pressure (PAP). **b**, Heart failure (HF) with reduced ejection fraction. **c**, Assessment of left ventricular hypertrophy at various degrees of severity (mild, moderate and severe). **d**, Left atrial dilation at various degrees of severity (mild, moderate and severe). **e**, Left ventricular dilation at various degrees of severity (mild, moderate and severe). **f**, Assessment of pericardial effusion size (small, moderate and large) as well as presence of tamponade physiology. Data are from the Cedars-Sinai Medical Center (CSMC; blue, $n = 100{,}994$) and Stanford Healthcare (SHC; red, $n = 5{,}000$). FPR, false positive rate; TPR, true positive rate.

PAP. EchoCLIP predicts LVEF on the held-out internal test dataset with a mean absolute error (MAE) of 8.4% and an MAE of 7.1% on an external test set of videos from the EchoNet-Dynamic dataset from Stanford Healthcare (Fig. 1). At key clinical LVEF thresholds, EchoCLIP achieves an AUC of 0.89–0.90 for an LVEF threshold of 50%, 0.93–0.94 for an LVEF threshold of 40% and 0.95–0.97 for an LVEF threshold of 30% (Supplementary Table 3 and Supplementary Fig. 1). Furthermore, EchoCLIP predicts estimated PAP with an MAE of 10.8 mm Hg on the internal test dataset and an MAE of 10.8 on the external test dataset (Fig. 2).

Mapping clinical text to echocardiogram images

Given the long length of an echocardiography report, we developed EchoCLIP-R, a domain-specific text encoder that succinctly summarized common cardiovascular concepts into fewer tokens and was able to summarize a whole report during training. A long-context EchoCLIP-R model was optimized for retrieval during training. Given a representative image from the held-out test cohort, EchoCLIP-R ranks the matching clinical report on average 209th out of 21,484 candidates (top 1% retrieval). The correct report is present in the top ten reports 33.3% of the time. Going from text to image, the average rank of the matching video is 203 out of 21,484 and the correct video is present in the top ten ranked videos 34.3% of the time. For all language models, the choice of text prompts impacts model performance and we found EchoCLIP to be easier to generate focused prompts compared to EchoCLIP-R given the larger context for in-domain prompts (Table 2). A workflow for automated preliminary assessment of echocardiogram studies by ensembling assessments across videos is shown in the Supplementary Video.

Detection of clinical differences between videos

The ability to measure the similarity between pairs of echocardiograms can also be used to identify a unique patient across multiple studies (a difficult task for human clinicians) as well as identify clinical changes over time. Comparing the cosine similarity between EchoCLIP-R embeddings of different echocardiography studies can help in challenging clinical scenarios. Pairs of EchoCLIP-R embeddings of echocardiograms are, on average, least similar if they come from two different patients (mean cosine similarity 0.40, 95% CI 0.39–0.41), more similar if they come from the same patient but were acquired on different dates (mean cosine similarity 0.64, 95% CI 0.64–0.65) and most similar if they come from the same patient and were acquired on the same day (mean cosine similarity 0.87, 95% CI 0.86–0.87). This comparison results in an AUC of 0.86 (95% CI 0.85–0.87) in identifying the same patients across different videos. Furthermore, the cosine similarity between videos can also be used to distinguish when there was a substantive clinical change. Echocardiograms acquired before cardiac surgeries and orthotopic heart transplants tend to be similar to one another, while being substantially less similar to echocardiograms acquired after such procedures (Fig. 3). This dropoff in embedding similarity is sufficient to predict whether an echocardiogram occurs before or after

https://doi.org/10.1038/s41591-024-02959-y

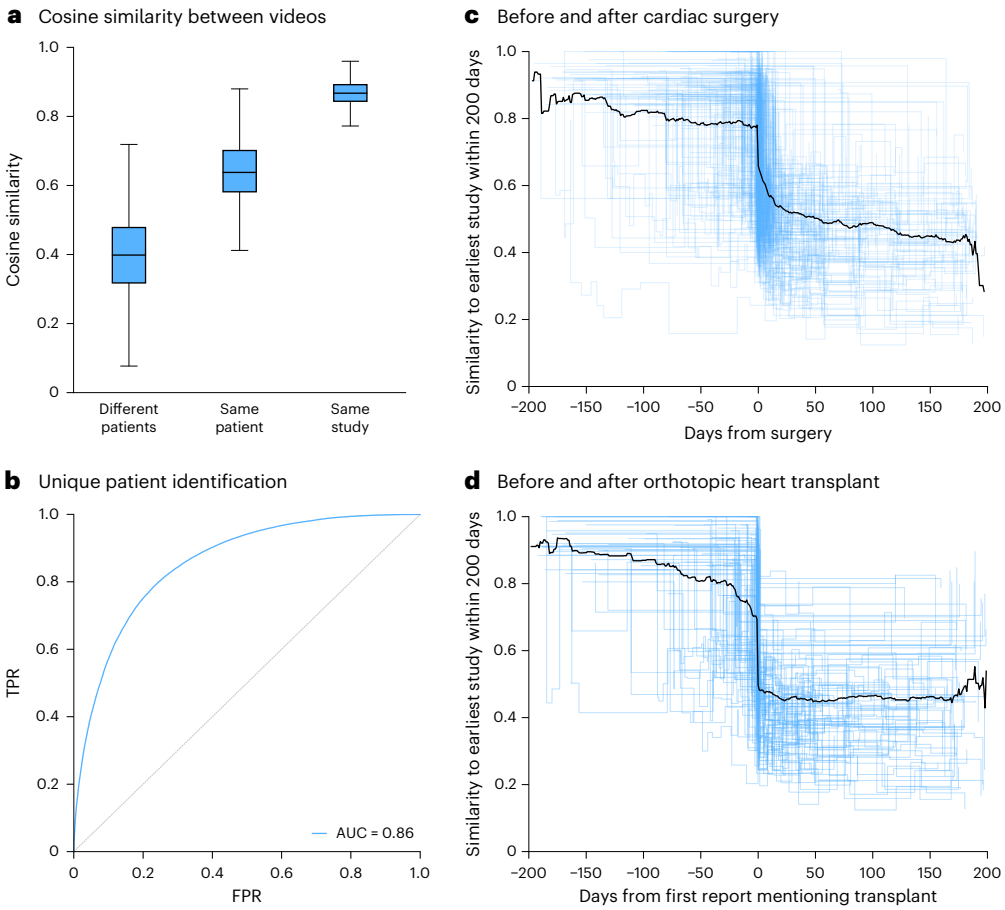


Fig. 3 | Assessment of clinical similarity. a, Average cosine similarity between embeddings from different patients, same patients at different times and same patients at the same time point. Center lines indicate the median, boxes span from the first to the third quartile and whiskers stretch 1.5 × the interquartile range (n = 100,994). **b**, AUC for predicting whether the images come from the same patient when compared to another image (n = 100,994). **c**,**d**, Trajectory of individual patients by cosine similarity (n = 2,959). Each line represents an individual patient with time from major clinical event on the x axis and cosine similarity versus first study on the y axis. Patients either had major cardiac surgery (**c**) or heart transplant (**d**), with cosine similarity calculated at the study level and pairwise compared for all videos in each study. Data are from the Cedars-Sinai Medical Center. FPR, false positive rate; TPR, true positive rate.

cardiac surgery with an AUC of 0.77 (95% CI 0.75–0.79) and before or after heart transplant with an AUC of 0.79 (95% CI 0.76–0.82). Additionally, we show that the difference in reported LVEF between different studies from the same patient is correlated with the cosine similarity between videos, suggesting that EchoCLIP-R embeddings can be used to identify clinically relevant serial changes (Supplementary Fig. 2).

Interpretation studies

To further interrogate EchoCLIP's understanding of cardiovascular disease, we utilized two interpretability frameworks. First, we developed a modified class activation mapping method (PromptCAM) for multimodal models that pairs textual prompts with imaging features. PromptCAM identifies regions of interest in the image that maximize the cosine similarity with the text prompts. Despite not seeking to minimize the loss of direct text labels in training, PromptCAM highlights the learned associations of EchoCLIP for subconcepts such as 'TAVR', 'Impella', 'Pacemaker' or 'Mitraclip' (Fig. 4). Secondarily, we applied Uniform Manifold Approximation and Projection (UMAP) on the embeddings from the EchoCLIP image encoder and observed numerous clusters associated with different cardiovascular diseases, disease states and measurements (Supplementary Fig. 3).

Discussion

Our results suggest that large datasets of echocardiography studies and expert adjudicated interpretations can serve as the basis for training medical foundation models. Our echocardiography foundation model was able to successfully complete multiple benchmarks of zero-shot prediction tasks without task-specific training or fine-tuning. By training EchoCLIP with data from one healthcare system and testing its performance on data from an entirely separate external healthcare system, we were able to evaluate EchoCLIP's generalizability and robustness to domain shift. Additionally, EchoCLIP-R displays an ability to perform tasks that human clinicians struggle with or find laborious, such as identifying the same patient across different imaging studies and characterizing clinically important changes over time. Finally, we introduce a multimodal interpretability approach using cosine similarity-based saliency to demonstrate that EchoCLIP has learned semantically meaningful imaging features of both common and rare cardiovascular concepts based on text prompting.

A key bottleneck in training medical foundation models is the limited availability of medical training data. Previous echocardiography AI models were trained with a maximum of 150,000 echocardiogram videos[29] and most frequently trained with only hundreds or thousands of examples[7,10,30–32]. By leveraging large clinical reporting databases, our approach minimizes the tedious manual labeling and organization required for supervised learning tasks and allows EchoCLIP to be trained on over 1 million echocardiography videos. In its most basic form, the task of mapping images to corresponding text interpretations is the clinical task of medical image interpretation that cardiologists do daily. EchoCLIP represents an opportunity to automate many

https://doi.org/10.1038/s41591-024-02959-y

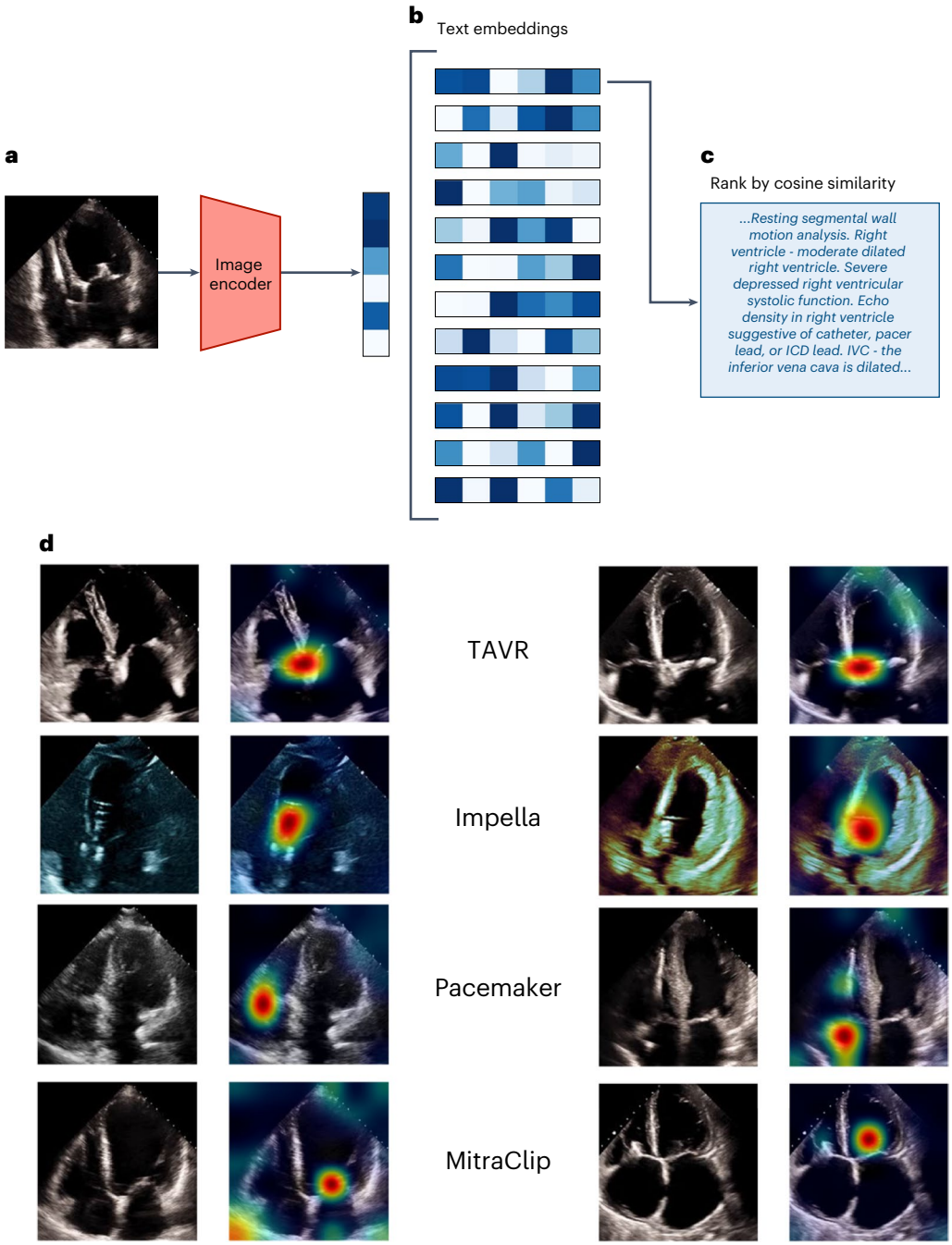


Fig. 4 | Image-to-text semantic search. a, The query image is first embedded using EchoCLIP-R's image encoder. **b**, Then, the similarities between this query embedding and the embeddings of all 21,484 unique text reports in the test set are computed. **c**, The reports are ranked by their similarity to the query image embedding and the report with the highest similarity is retrieved. **d**, Corresponding pairs of input frames and PromptCAM visualization of the indicated intracardiac devices in the text report label (color intensity ranging from red for most important to green for less important and no color for not important).

interpretation tasks simultaneously without the need for individually tuned specialist models, which can ultimately lead to automated preliminary echocardiography interpretation in underserved populations or during emergent situations. A self-supervised vision–language foundation model trained on the diverse range of physiologies seen in a high-volume echocardiography laboratory can learn from greater amounts of data than purely supervised models and may be able to gain a much more generally applicable understanding of the human heart, its function and its structure.

While EchoCLIP is not the first instance of a foundation model trained on biomedical datasets[17,33,34], EchoCLIP is a model specific for echocardiography, the most common modality for cardiovascular imaging and not represented in prior foundation model training. While echocardiography still needs to be interpreted by expert cardiologists, given the rapid expansion of availability of ultrasound technology and the development of complementary technologies to allow novices to perform cardiac ultrasound[35], models such as EchoCLIP have the potential to improve access to cardiac imaging and image interpretation. One of the most time-consuming and challenging assessments is distinguishing between natural variation versus change in the disease state that might warrant changes in the treatment plan. Such evaluations often require meticulously comparing current and historical imaging side by side and can be highly variable across different cardiologists. By using EchoCLIP to directly compare studies, clinicians can derive a

quantitative visual assessment of differences. Such an automated AI assessment can alert clinicians' attention toward specific studies to more carefully evaluate clinical changes.

While specialist models still perform better on specific, narrowly defined tasks[29], the performance of EchoCLIP on external validation data confirms its ability to assess cardiac function with accuracy similar to blinded human performance as well as many previously developed supervised learning models[6,7,36–38]. EchoCLIP achieves an MAE of 7.1% on external validation of LVEF prediction, while previous video-based LVEF AI models achieve an MAE of 6.0% and image-based AI models achieve an MAE of just 9.9%[28]. Differences in EchoCLIP's performance on internal and external test datasets are likely due to differences in the LVEF evaluation technique across institutions (Supplementary Fig. 4). While statistically significant in large datasets, the error in model predictions across LVEF values or measurement approaches is less than clinical variability[29], suggesting that different training data with a different LVEF measurement approach would have a modest differential effect. The distribution of LVEF values from model inference is continuous without preference for certain measurements, suggesting that human biases are smoothed out in the model embedding space (Supplementary Fig. 5).

Important limitations of this work include the use of an image encoder instead of a video encoder when echocardiography videos contain important motion-based information and the use of only the apical-four-chamber view, which, although is the most common and informative standard view, does not capture information with regard to Doppler velocities and structures only present in other views. In this work, as well as previous work[8,9], it is clear that there are image-based features that can be a partial surrogate for information not directly interrogatable without video or from different views. For example, sphericity and dilation of the left ventricle can be identifiable from images alone and suggest decreased cardiac function although true assessment of LVEF requires video information. Valve calcification can hint at stenosis or coronary artery disease[8,9,39] that is not directly present in the image. Future work will incorporate video encoders and different measurement techniques and will leverage multiple views from the same echocardiographic study to provide more holistic AI models for heart health. Enhancements such as upgrading EchoCLIP's visual encoder from an image-based model to a video-based model, adapting EchoCLIP for visual question answering, and implementation of automatic report generation are potential directions for future research. Finally, important open questions remain in the testing of foundation models before regulatory approval and eventual clinical use.

Our results encourage further exploration of vision–language foundation models for cardiology and medicine generally. Clinical databases provide large bodies of information about health, while different imaging modalities provide adjunctive ancillary information that might improve our understanding of cardiovascular health. Further efforts remain to leverage larger datasets and more versatile model architectures to better capture and distill medical information.

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/s41591-024-02959-y.

References

1. Heidenreich, P. A. et al. 2022 AHA/ACC/HFSA guideline for the management of heart failure: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* **145**, e876–e894 (2022).
2. Al-Khatib, S. M. et al. 2017 AHA/ACC/HRS guideline for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Rhythm Society. *Circulation* **138**, e210–e271 (2018).
3. Wilcox, J. E., Fang, J. C., Margulies, K. B. & Mann, D. L. Heart failure with recovered left ventricular ejection fraction: JACC Scientific Expert Panel. *J. Am. Coll. Cardiol.* **76**, 719–734 (2020).
4. Dunlay, S. M., Roger, V. L. & Redfield, M. M. Epidemiology of heart failure with preserved ejection fraction. *Nat. Rev. Cardiol.* **14**, 591–602 (2017).
5. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
6. Zhang, J. et al. Fully automated echocardiogram interpretation in clinical practice. *Circulation* **138**, 1623–1635 (2018).
7. Tromp, J. et al. Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. *Lancet Digit. Health* **4**, e46–e54 (2022).
8. Holste, G. et al. Severe aortic stenosis detection by deep learning applied to echocardiography. *Eur. Heart J.* **44**, 4592–4604 (2023).
9. Ghorbani, A. et al. Deep learning interpretation of echocardiograms. *NPJ Digit. Med.* **3**, 10 (2020).
10. Duffy, G. et al. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA Cardiol.* **7**, 386–395 (2022).
11. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at https://arxiv.org/abs/2108.07258 (2021).
12. Radford, A. et al. Learning transferable visual models from natural language supervision. in *Proc. 38th International Conference on Machine Learning* Vol. 139 (PMLR, 2021).
13. Desai, K. & Johnson, J. VirTex: learning visual representations from textual annotations. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2021).
14. Larochelle, H., Erhan, D. & Bengio, Y. Zero-data learning of new tasks. in *Proc. 23rd AAAI Conference on Artificial Intelligence* (AAAI, 2008).
15. Geirhos, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. Preprint at https://arxiv.org/abs/1811.12231 (2018).
16. Eslami, S., de Melo, G. & Meinel, C. Does CLIP benefit visual Question answering in the medical domain as much as it does in the general domain? Preprint at https://arxiv.org/abs/2112.13906 (2021).
17. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
18. Ji, S. et al. Domain-specific continued pretraining of language models for capturing long context in mental health. Preprint at https://arxiv.org/abs/2304.10447 (2023).
19. Thawkar, O. et al. XrayGPT: chest radiographs summarization using medical vision-language models. Preprint at https://arxiv.org/abs/2306.07971 (2023).
20. Iyer, N. S. et al. Self-supervised pretraining enables high-performance chest X-ray interpretation across clinical distributions. Preprint at *medRxiv* https://doi.org/10.1101/2022.11.19.22282519 (2022).
21. Liu, Z. et al. Radiology-GPT: a large language model for radiology. Preprint at https://arxiv.org/abs/2306.08666 (2023).
22. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual-language foundation model for pathology image analysis using medical Twitter. *Nat. Med.* **29**, 2307–2316 (2023).
23. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
24. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).

25. Abbaspourazad, S. et al. Large-scale training of foundation models for wearable biosignals. Preprint at https://arxiv.org/abs/2312.05409 (2023).
26. Liu, Z. et al. A ConvNet for the 2020s. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2022).
27. Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units. in *Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, 2016).
28. Cherti, M. et al. Reproducible scaling laws for contrastive language-image learning. Preprint at https://arxiv.org/abs/2212.07143 (2022).
29. He, B. et al. Blinded, randomized trial of sonographer versus AI cardiac function assessment. *Nature* **616**, 520–524 (2023).
30. Lau, E. S. et al. Deep learning-enabled assessment of left heart structure and function predicts cardiovascular outcomes. *J. Am. Coll. Cardiol.* **82**, 1936–1948 (2023).
31. Akerman, A. P. et al. Automated echocardiographic detection of heart failure with preserved ejection fraction using artificial intelligence. *JACC Adv.* **2**, 100452 (2023).
32. Madani, A., Arnaout, R., Mofrad, M. & Arnaout, R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit. Med.* **1**, 6 (2018).
33. Lee, J. et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
34. Rasmy, L., Xiang, Y., Xie, Z., Tao, C. & Zhi, D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit. Med.* **4**, 86 (2021).
35. Narang, A. et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. *JAMA Cardiol.* **6**, 624–632 (2021).
36. Farsalinos, K. E. et al. Head-to-head comparison of global longitudinal strain measurements among nine different vendors: the EACVI/ASE inter-vendor comparison study. *J. Am. Soc. Echocardiogr.* **28**, 1171–1181 (2015).
37. Yuan, N. et al. Systematic quantification of sources of variation in ejection fraction calculation using deep learning. *JACC Cardiovasc. Imaging* **14**, 2260–2262 (2021).
38. Cole, G. D. et al. Defining the real-world reproducibility of visual grading of left ventricular function and visual estimation of left ventricular ejection fraction: impact of image quality, experience and accreditation. *Int. J. Cardiovasc. Imaging* **31**, 1303–1314 (2015).
39. Yuan, N. et al. Prediction of coronary artery calcium using deep learning of echocardiograms. *J. Am. Soc. Echocardiogr*. https://doi.org/10.1016/j.echo.2022.12.014 (2022).

Methods

Data curation

The Cedars-Sinai Medical Center echocardiography laboratory performs clinical echocardiography for a wide range of indications, ranging from asymptomatic preoperative screening to evaluation for open heart surgery or heart transplant. Over the course of a standard, full, resting echocardiogram study, 50–150 videos and images are acquired that visualize the heart from different angles, locations and with different imaging modes (two-dimensional images, tissue Doppler images and color Doppler images). Each echocardiogram study corresponds to a unique patient and a unique visit, but multiple similar videos may be obtained from each view acquired during the study. For EchoCLIP, we focused on the apical-four-chamber view (one of the most common and well-acquired ultrasound views) and organized a dataset of 1,032,975 unique video–caption pairs from 224,685 echocardiogram studies across 99,870 patients, collected between 2011 and 2022. Our laboratory developed high-throughput tools to query echocardiogram videos and their metadata from Cedars-Sinai's internal databases at scale, view and classify videos and link them to associated structured reporting from cardiologists. DICOM images were queried from a Hyland vendor-neutral archive, linked to interpretations created by trained cardiologists using Syngo Dynamics and converted to AVI video files using PyDICOM before model training and inference.

Data were split by patient into training, validation and internal test datasets. The training data contained 921,981 videos from 84,990 patients, the validation set contained 10,000 videos from 5,358 patients and the internal test set contained 100,994 videos from 10,001 patients. A random subset ($n = 5{,}000$) of the publicly released EchoNet-Dynamic dataset from Stanford Healthcare was used as an external test set. An automated preprocessing workflow was undertaken to remove extraneous text, ECG and respirometer information and other information outside of the scanning sector. The input data were represented as standardized 224 × 224-pixel RGB videos for model training. This research was approved by the Cedars-Sinai Medical Center (study no. 00001409) and Stanford Healthcare Institutional Review Boards (study no. 43721). A waiver of consent was obtained for the use of retrospective de-identified data.

Model design and training

Model design and training was conducted in Python using the PyTorch deep-learning library. Our training code is a fork of the OpenCLIP repository[28]. To find the best training configuration, we evaluated a variety of model architectures and training procedures. We tested training with random initialization, initializing the model with CLIP weights, using a convolutional architecture for the image encoder, using a vision transformer for the image encoder, applying random patch dropout to image inputs and using three different text tokenization methods (Supplementary Table 5), with the final EchoCLIP model use the ConvNeXt architecture[26] for the image encoder and a decoder-only transformer for the text encoder. We initialize our model with weights pretrained on LAION-400M. We trained for 50 epochs, minimizing the original CLIP loss. The CLIP loss incentivizes the video and text encoders to make the embeddings of paired videos and reports as similar as possible, while making the embeddings of unpaired videos and reports as different as possible (Fig. 1a). This training objective is, notably, all that is required to make the two models learn to encode their inputs into semantically meaningful vector embeddings.

We warmed up to an initial learning rate of 5×10^{-5} over the course of the first 2,000 training steps and then cosine decayed to zero over the course of the training run. We used a batch size of 1,024 and trained on two Nvidia RTX A6000 48 GB GPUs for approximately 2 weeks. During training, a random frame was extracted from each video and passed to the image encoder. A random frame from each video was used for each epoch as a form of data augmentation. Model checkpoints were saved after every epoch. At the end of training, the model checkpoint with the lowest mean cross-modal retrieval rank on the validation set was selected for testing. Before computing the cosine similarity between vector embeddings, we always divide them by their norms to ensure that they have the same magnitude. This means that the cosine similarity metric always returns a value between −1 and 1.

Text tokenization

A number of text tokenization schemes were tested (Supplementary Table 1). EchoCLIP was trained using text tokenized by a BPE tokenizer[27] pretrained on the GPT2 data corpus, which encoded echocardiography reports with a mean of 530.3 (±154.7) tokens per report. Due to the context length limit of 77 tokens imposed by fine-tuning from CLIP weights, EchoCLIP was trained on snippets of reports rather than their full text. For EchoCLIP-R, we noted that the echocardiography report text is often highly structured and repetitive, as they are typically generated in a 'fill-in-the-blank' fashion according to a predetermined template given to the cardiologist at the time of interpretation. The templated nature of the reports means that a small number of unique phrases and sentences appear very frequently in the final report text with only slight variations. A custom-built tokenizer was designed to take advantage of this observation and allowed us to aggressively compress the report text. This meant that whole reports could be inputted when training EchoCLIP-R, improving its retrieval capabilities compared to EchoCLIP at the cost of slight degradation in classification and estimation capabilities.

Instead of searching for exact vocabulary matches in the report text, our custom-built template tokenizer uses regular expressions to allow nearly similar lines of text to be efficiently encoded. For example, the text 'Moderate left ventricular hypertrophy. Left ventricular ejection fraction is 60%' is converted into tokens indicating either cardiac structure or function (such as '<_left ventricular hypertrophy>', '<left ventricular ejection fraction is _%>') as well as indicating severity ('mild', 'moderate' or 'severe') or quantity (60%, 2.5 cm, 40 cm s^{-1}). By doing this, we were able to capture most of the variance present in our text reports with a vocabulary containing only 770 words and phrases, in addition to extra tokens for handling numbers and severity terms. After applying this custom tokenizer, the mean length of a tokenized report was brought down to just 63.8 (±26.7) tokens, an approximate ninefold reduction compared to using CLIP's original BPE tokenizer. We additionally tested a model that used a BPE tokenizer pretrained on echocardiography reports but found that it failed to outperform the model trained using our custom solution.

Using EchoCLIP-R embeddings, we can perform a search within our test set to find images or reports that are semantically similar to a given query image or report. To do this, we simply sort the embeddings of all candidate images or reports by their cosine similarity to the embedding of a query image or report. The embedding space was normalized to unit vectors before calculation of cosine similarity to be insensitive to projection magnitude. If the model and dataset were theoretically perfect, we would expect the image or report that is officially paired with the query image or report to be ranked first in the list. We report the mean rank number as a metric of accuracy. This allows us to characterize EchoCLIP-R's retrieval abilities in two settings: image-to-report and report-to-image. We choose a single random video from each study to represent the whole study in these ranking tests to simplify the implementation. To obtain a single value that represents a model's overall retrieval ability, we define the MCMRR as the average of both the mean image-to-report retrieval rank and the mean report-to-image retrieval rank. MCMRR values for both EchoCLIP and EchoCLIP-R are shown in Table 2.

To evaluate the model's ability to identify unique patients, we computed the similarity between many random pairs of EchoCLIP-R's image embeddings and then treated those similarity values as if they were continuous probability predictions meant to classify whether both images in the pair came from the same patient. To visualize patient

trajectories before or after heart transplantation or cardiac surgery, we first collected all the echocardiogram images within 200 days before or after the procedure date. These images were grouped by study and then embeddings were produced for each video using EchoCLIP-R. The earliest study within the 200-day window was taken as a baseline and then each following study within the window was assigned a similarity score computed by taking the average similarity between all possible pairs of videos from the baseline study and the study in question. This was repeated for all patients who had undergone heart transplantation or cardiac surgery in our test set who also had at least one echocardiography study performed before and after the date of the procedure. These study-level 'similarity timelines' were then plotted together, resampled and averaged to create Fig. 3c,d. These study-level similarity scores can also treated as continuous probability predictions for whether a given study was acquired before or after the procedure date, allowing us to calculate an AUC score that quantifies EchoCLIP-R's ability to detect the effects of such procedures. Multiple surgical characteristics and approaches were analyzed by subset analysis with similar results (Supplementary Figs. 6 and 7).

Adapting EchoCLIP to classification and regression tasks

Despite only training to encode images and report text as semantically meaningful vector embeddings, EchoCLIP was adapted to perform both classification and regression tasks. For each classification task, we followed the approach of the original CLIP paper and constructed text prompts describing a positive case. Then, we obtained an embedding of those prompts using EchoCLIP's text encoder and computed the cosine similarity between them and the embeddings of the videos in our test set. In the case of multiple semantically equivalent prompts being used for a binary classification task, we average the similarity across all prompts and then averaged again over the first 20 frames of the video (at temporal stride 2). We treat this final average similarity score as a continuous probability prediction. Hyperparameters of number of frames sampled per video, stride (frame count between sampled frames) and number of averaged embeddings were evaluated to optimize model performance (Supplementary Tables 4 and 5).

For regression tasks, we generated a collection of variations on a base text prompt by only changing the relevant value in the text (Supplementary Fig. 8). For instance, variants of the prompt 'The left ventricular ejection fraction is estimated to be X%' or 'LV ejection fraction is X%' were generated for all integer values between 0 and 100. These variations on the base prompt are then embedded using EchoCLIP's text encoder. The cosine similarity between these prompt embeddings and the embeddings of each of the first 20 frames of all test-set videos (extracted with temporal stride of 2) is computed. The candidate values are then ranked for each frame according to their corresponding prompt embeddings' similarity to the frame embeddings and the bottom 80% of the values are discarded. The remaining 20% of the values are averaged along the frames dimension, leaving 20 potential prediction values ordered from most likely (on average across all frames) to least likely. We found, empirically, that taking the median of these 20 values results in the most accurate predictions. This process is illustrated in Extended Data Fig. 1.

For EchoCLIP, a systematic search through relevant phrases present in the echocardiography report template file was conducted to manually construct the base prompts for each task. For EchoCLIP-R, we noted that using this approach resulted in severely degraded performance. We believe this to be the result of short, single-phrase prompts being out-of-distribution for EchoCLIP-R as it was trained exclusively using full-length reports. To address this, we tested an alternate prompting strategy for EchoCLIP-R, where the base prompts are entire reports sampled from videos in the validation set that have the desired labels. As an example of how this works for a regression task, the base LVEF estimation prompts for EchoCLIP-R were chosen by randomly sampling up to ten reports from the validation set for each ground-truth LVEF value between 1 and 100. This way, EchoCLIP-R has in-distribution 'example reports' from the validation set to compare the query images against, instead of being forced to encode much shorter prompts that are nothing like what it saw during training. For binary tasks, 200 reports containing a positive label for the task are sampled from the validation set and used as base prompts. We found that this 'sampled prompts' strategy substantially improved EchoCLIP-R's performance on classification and regression tasks (Table 2).

All text prompts used for the evaluation of EchoCLIP are published in the project's code repository, a link to which is included in Supplementary Fig. 8. Ground-truth labels are extracted from the clinical reports and used to calculate AUC and other performance metrics.

Interpretation techniques

Code for saliency mapping with PromptCAM was written in Python with dependencies on PyTorch and NumPy packages. Modifying the optimization function of the integrated gradients method, PromptCAM maximizes the cosine similarity as the objective function between image-based regions of interest with the text prompt. Prompts describing common cardiac structures were used to test whether EchoCLIP 'pays attention' to relevant cardiac structures in echocardiogram images. UMAP was applied using the umap-learn Python package. EchoCLIP image embeddings for each video in the test set were processed to demonstrate how clusters associated with different cardiovascular diseases, disease states and measurements are present. The n_neighbors parameter was set to the maximum allowed value of 200 and the min_distance parameter was set to the maximum allowed value of 1.0.

Reporting summary

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

Data availability

The dataset of videos and reports used to train EchoCLIP is not publicly available due to its potentially identifiable nature; however, EchoNet-Dynamic, the dataset that we used for external validation, is publicly available at https://echonet.github.io/dynamic/.

Code availability

Our model weights, evaluation prompts, and demonstration code are available on GitHub at https://github.com/echonet/echo_CLIP.

Acknowledgements

We thank B. He and G. Duffy for thoughtful discussions and feedback. D.O. discloses National Institutes of Health NHLBI grants R00HL157421 and R01HL173526.

Author contributions

M.C. developed the model. M.C., M.V. and N.Y. performed the experiments. M.C. and D.O. wrote the paper. All authors provided critical feedback and review.

Competing interests

The authors declare no competing interests.

Additional information

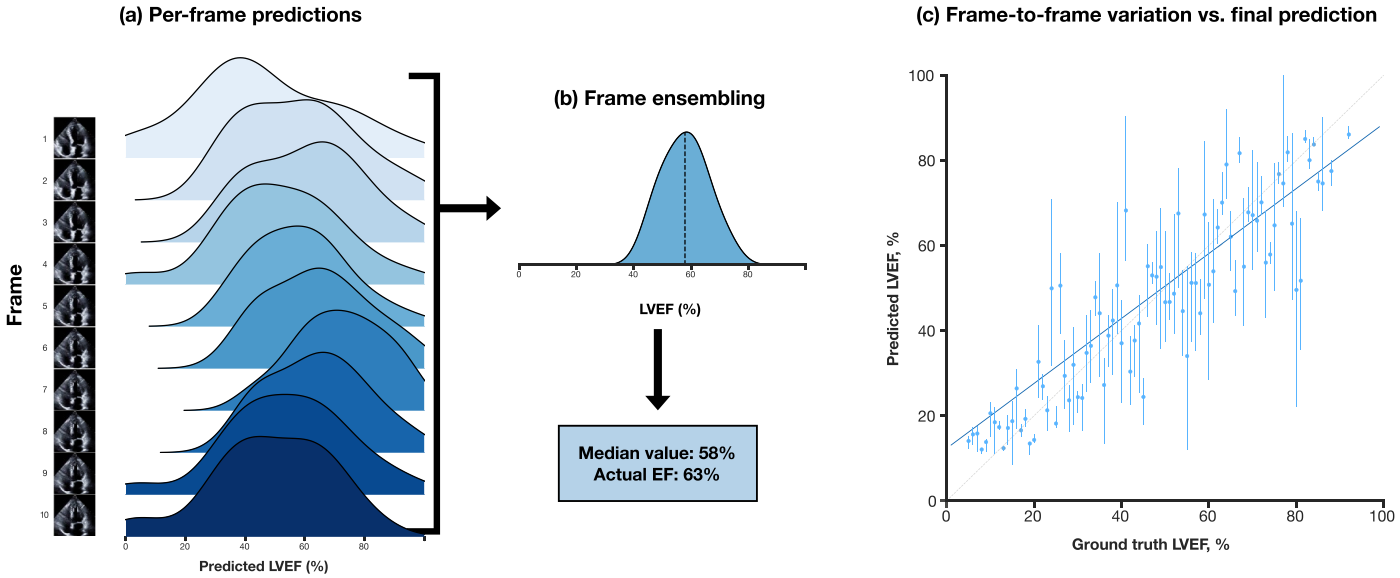
Extended data is available for this paper at https://doi.org/10.1038/s41591-024-02959-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-02959-y.

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

Peer review information *Nature Medicine* thanks Darrel Francis, Massoud Zolgharni and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Michael Basson, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-024-02959-y



Extended Data Fig. 1 | Frame level ensembling. (**a**) Distribution of EchoCLIP left ventricular ejection fraction (LVEF) from individual frames of an echocardiogram video, which are averaged to (**b**) a video-level distribution of LVEF prediction. (**c**) Scatter-plot of subset of test dataset (n = 1,000 predictions from 100 videos and 10 frames per video) representing predicted vs. ground-truth LVEF. Each point represents the final predicted values and whiskers represent the range of frame level predictions for that video.

https://doi.org/10.1038/s41591-024-02959-y

Extended Data Table 1 | AUC, sensitivity, and specificity for EchoCLIP zero-shot prediction tasks

Zero-Shot Task	AUC	Sensitivity	Specificity
Severe left atrial dilation	0.91	0.89	0.77
Moderate left atrial dilation	0.85	0.86	0.68
Mild left atrial dilation	0.74	0.74	0.62
Severe left ventricular dilation	0.92	0.81	0.87
Moderate left ventricular dilation	0.86	0.89	0.68
Mild left ventricular dilation	0.78	0.78	0.65
Presence of MitraClip	0.97	0.89	0.94
Presence of Transcatheter Aortic Valve Replacement	0.92	0.83	0.86
Presence of Pacemaker	0.84	0.70	0.82
Left Ventricular Ejection Fraction Below 50% (CSMC)	0.90	0.86	0.79
Left Ventricular Ejection Fraction Below 50% (SHC)	0.89	0.86	0.79
Severe Left Ventricular Hypertrophy	0.82	0.64	0.80
Moderate Left Ventricular Hypertrophy	0.75	0.71	0.63
Mild Left Ventricular Hypertrophy	0.65	0.80	0.39
Tamponade	0.96	0.86	0.95
Large Pericardial Effusion	0.96	0.94	0.86
Moderate Pericardial Effusion	0.94	0.85	0.90
Small Pericardial Effusion	0.71	0.51	0.79

Area under the receiver operator curve, sensitivity, and specificity for zero-shot classification tasks. Sensitivity and specificity calculated at the Youden's index.

Area under the receiver operator curve, sensitivity, and specificity for zero-shot classification tasks. Sensitivity and specificity calculated at the Youden's index.

nature portfolio

Corresponding author(s): Christensen et al

Last updated by author(s): 3/26/2024

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. Medical images were converted from initial DICOM files into AVI videos files prior to deep learning training. Associated text and patient identifiers were mapped from the electronic healthcare record for training but videos we de-identified prior to input into AI model.
Data analysis	A deep learning algorithm was used to assess the echocardiogram videos. Zero shot tasks were performed with text prompts. Our code and working model is available at https://github.com/echonet/echo-clip. Required software packages for model training and inference include pytorch, torchvision, open_clip_torch, huggingface_hub, tokenizers, opencv-python-headless, pathlib, numpy, re, matplotlib, os, PIL, and scipy.

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

The dataset of videos and reports used to train EchoCLIP is not publicly available due to its potentially identifiable nature. However, EchoNet-Dynamic, the dataset we used for external validation, is publicly available at https://echonet.github.io/dynamic/.

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. Model performance in sex stratified populations shown in Supplementary Table 2.
Reporting on race, ethnicity, or other socially relevant groupings	The input data for training comes from a large academic medical center with a 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. External validation data from Ouyang et al. Nature 2020 publicly released data and cohort demographics in the prior manuscript.
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) 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	1,032,975 echocardiograms were collected from 224,685 unique studies performed on 99,870 unique 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 off of availability of echocardiogram videos for training at the healthcare site. Given prior echocardiogram AI models are trained on much less data (10-100x smaller training dataset sizes), we anticipated use of greater than 1 million samples would be sufficient to train a foundation model.
Data exclusions	Echocardiograms not classified as apical-4-chamber were excluded from the study.
Replication	95% confidence intervals were calculated using bootstrapping. The algorithm otherwise is deterministic and code is available.
Randomization	Patients were randomly divided into training, validation, and testing splits with approximate ratio of 89:1:10.
Blinding	This study is wholly retrospective and no additional human input was collected. Blinding was therefore neither possible nor necessary for this study.

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

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Not a clinical trial
Study protocol	No prospective human subject activities were undertaken. Model design and training is described in the manuscript.
Data collection	The data was collected from the picture archival and communication system (PACS). This research was IRB approved.
Outcomes	The data was collected from the electronic healthcare record and structured reporting system. This research was IRB approved.

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.*