Title: Comparative Eminence: Foundation versus Domain-Specific Model for Cardiac Ultrasound Segmentation

Authors: Chieh-Ju Chao, MD[1,2], Yunqi Richard Gu[2], Tiange Xiang, BS[2], Lalith Appari, BS[3,4], Justin Wu, BS[2], Juan M. Farina, MD[5], Rachael Wraith, RDCS[5], Jiwoon Jeong, MS[3,4], Reza Arsanjani, MD[5], Garvan C. Kane, MD[1], Jae K. Oh, MD[1], Curtis P. Langlotz, MD, PhD[2], Imon Banerjee, PhD[3,4] , Li Fei-Fei[†], Ph.D.[2], Ehsan Adeli[†], Ph.D.[2]

[†]Li Fei-Fei and Ehsan Adeli are co-corresponding authors on this manuscript.

Affiliations:

1. Department of Cardiovascular Medicine, Mayo Clinic Rochester, Rochester, MN
2. Stanford Institute for Human-Centeral Artificial Intelligence, Stanford University, Stanford, CA
3. Department of Radiology, Mayo Clinic Arizona, Scottsdale, AZ
4. School of Computing and Augmented Intelligence, Arizona State University, Phoenix, AZ
5. Department of Cardiovascular Medicine, Mayo Clinic Arizona, Scottsdale, AZ

Word count: 2,836 words in the manuscript body; 3 Figures, 2 Tables, 39 References. 3 Supplemental Tables.

Abbreviations:
AI: Artificial intelligence
A2C: apical 2 chamber, echocardiography view
A4C: apical 4 chamber, echocardiography view
CAMUS: Cardiac Acquisitions for Multi-structure Ultrasound Segmentation
IoU: Intersection over the union
DSC: Dice similarity coefficient
LV: Left ventricle
POCUS: Point-of-care ultrasound
SAM: Segment anything model
TTE: Transthoracic echocardiography
Vision Transformer: ViT

Corresponding Author:
Ehsan Adeli, Ph.D., Clinical Assistant Professor
Department of Psych/Major Laboratories and Clinical & Translational Neurosciences Incubator
Email: eadeli@stanford.edu
Address: MC#5717, 1070 Arastradero Rd., Palo Alto, CA 94304

Key points:

Question: What is the comparative performance of fine-tuned Segment Anything Model (SAM) against domain-specific segmentation model on transthoracic echocardiography (TTE) and point-of-care ultrasound (POCUS)?

Findings: Fine-tuned SAM had excellent performance on EchoNet dataset (SAM vs. EchoNet: DSC 0.911 ± 0.045 vs. 0.915 ± 0.047, p<0.0001) and generalized well on external datasets containing TTE (Mayo TTE: DSC 0.902 ± 0.032 vs. 0.893 ± 0.090, p<0.0001) and POCUS (DSC 0.857 ± 0.047 vs. 0.667 ± 0.279, p<0.0001).

Meaning: The generalization capability of SAM can facilitate the development of AI applications in echocardiography and POCUS with minimal expert data curation.

Abstract

Importance A recently developed vision foundation model, "Segment Anything (SAM)," promises to segment any objects in images. However, the performance of SAM on clinical echocardiography images is yet to be investigated and compared against the domain-specific models.

Objective To evaluate the performance of SAM on transthoracic echocardiography (TTE) and point-of-care ultrasound (POCUS) images.

Design SAM was fine-tuned on the training set of EchoNet-Dynamic (TTE) and then evaluated on datasets containing TTE and POCUS images.

Setting Multi-center, retrospective cohort study.

Participants This study used two publicly available datasets (EchoNet-dynamic, Stanford University and CAMUS, University Hospital of St Etienne). The Mayo Clinic dataset contains a sample of 99 non-duplicated patients (58 TTE and 41 POCUS).

Intervention/Exposure: not applicable.

Main Outcomes and Measures Model segmentation performance: Dice similarity coefficient (DSC).

Results Fine-tuned SAM had promising frame-level performance (SAM vs. EchoNet: DSC 0.911 ± 0.045 vs. 0.915 ± 0.047, p<0.0001), and consistent performance on the external datasets including TTE (Mayo Clinic: DSC 0.902 ± 0.032 vs. 0.893 ± 0.090, p<0.0001, CAMUS-A4C: DSC 0.897 ± 0.036 vs. 0.850 ± 0.097, p<0.0001, CAMUS-A2C: DSC 0.891 ± 0.040 vs. 0.752 ± 0.196, p<0.0001) and POCUS (DSC 0.857 ± 0.047 vs. 0.667 ± 0.279, p<0.0001).

Conclusions and Relevance Promising segmentation performance was observed after fine-tuning the SAM model on TTE. The strong generalization capability of SAM can facilitate the development of AI applications in cardiac ultrasound with less manual data curation.

Introduction

Echocardiography provides comprehensive anatomy and physiology assessment of the heart and is one of the most widely available imaging modalities in the field of Cardiology given its non-radiative, safe, and low-cost nature[1–3]. Cardiac chamber quantification is one of the fundamental tasks of echocardiography studies in the current practice[4], and the results can have direct effects on clinical decisions such as the management of heart failure, valvular heart diseases, and chemotherapy-induced cardiomyopathy [5–8]. Although left ventricular (LV) chamber quantification tasks are performed by trained sonographers or physicians, it is known to be subject to intra- and inter-observer variance which can be up to 7%-13% across studies[9–12]. Many artificial intelligence (AI) applications have been applied to address this essential task and to minimize variations[3,13–16].

While established AI systems are potential solutions, the training of segmentation AI models requires large amounts of training data and their corresponding expert-defined annotations, making them challenging and expensive to implement[17]. In recent years, transformers, a type of neural network architecture with self-attention, have revolutionized the field from natural language processing to computer vision[18–20]. Vision transformers (ViT) [18] are a type of transformer specifically designed for images, which have shown impressive performance with simple image patches and have become a popular choice for the building of foundation models that can be fine;-tuned for various downstream computer vision tasks[21]. Building on this success, Meta AI introduced the "Segment Anything Model" (SAM), a foundation large vision model that was trained on diverse datasets and that can adapt to specific tasks. This model achieves "zero-shot" segmentation: segments user-specified objects at different data resources without needing any training data[22].

However, while the zero-shot performance of SAM on natural image datasets has been promising[22], its performance on complex image datasets, such as medical images, has not been fully investigated. While not specifically including echocardiography images, one study tested SAM on different medical image datasets including ultrasound, and the zero-shot performance was not optimal[23]. Recently, MedSAM was introduced as a universal tool for medical image segmentation, however, ultrasound or echocardiography were less represented in the training set[24]. In this context, we aim to study the zero-shot and fine-tuned performance of SAM in echocardiography images and represent a comparative performance with a state-of-the-art segmentation model trained with a domain-specific dataset (EchoNet).

We hypothesized that SAM has suboptimal zero-shot segmentation performance on echocardiography images given the domain differences and the datasets' complexity. We also anticipate that fine-tuning SAM can adapt it to the domain of echocardiography for better segmentation performance. Within the echocardiography domain, images obtained from different institutions and modalities (with different image qualities) were used to evaluate the generalization capability of SAM performance.

Method

Population and Data Curation

The EchoNet-Dynamic dataset is publicly available (https://echonet.github.io/dynamic/); details of the

dataset have been described previously [3]. In brief, the dataset contains 10,030 apical-4-chamber (A4C) TTE videos at Stanford Health Care in the period of 2016-2018. The raw videos were preprocessed to remove patient identifiers and downsampled by cubic interpolation into standardized 112 × 112-pixel videos. Videos were randomly split into 7,465, 1,277, and 1,288 patients, respectively, for the training, validation, and test sets[3]. In this study, the cases without ground truth labels were excluded from this analysis (5 from the train set, 1 from the test set)(**Supplemental Table 1**).

A dataset from Mayo Clinic (Rochester, MN) that includes 99 randomly selected patients (58 TTE in 2017-2018 and 41 point-of-care ultrasound studies (POCUS) in 2022) was used as the external validation dataset. The A4C videos were reviewed by a clinical sonographer (RW) and a cardiologist (JMF). 52 (33 TTE and 19 POCUS) out of the total 100 cases were traced by both of the annotators to select and segment the end-diastolic and end-systolic frames. The tracings were done manually on commercially available software (MD.ai, Inc., NY). The Cardiac Acquisitions for Multi-structure Ultrasound Segmentation (CAMUS) dataset contains 500 cases from the University Hospital of St Etienne (France) with detailed tracings on both A4C and A2C views [25].

Segment Anything Model (SAM)

The Segment Anything Model (SAM) is an image segmentation foundation model trained on a dataset of 11 million images and 1.1 billion masks[26]. It can generate object masks from input prompts like points or boxes. SAM's promptable design enables zero-shot transfer to new image distributions and tasks, achieving competitive or superior performance compared to fully supervised methods[26]. In brief, the model comprises a VisionEncoder, PromptEncoder, MaskDecoder, and Neck module, which collectively process image embeddings, point embeddings, and contextualized masks to predict accurate segmentation masks.

Data Preprocessing

Each EchoNet-Dynamic video (112 × 112 pixels, in avi format) was exported into individual frames without further resizing. End-diastolic and end-systolic frames of each case were extracted, which corresponded to the human expert traced, frame-level ground truth segmentation coordinates in the dataset. Ground truth segmentation masks were generated according to the coordinates and saved in the same size (112 × 112 pixels). The labeled frames of Mayo TTE and POCUS images were exported from the MD.ai platform, and followed a similar preprocessing method to remove patient identifiers, then horizontally flipped and resized to 112 × 112 pixels[3]. The CAMUS images were rotated for 270 degrees and resized to be consistent with the EchoNet format. When importing to the SAM model, all the raw images were resized with the built-in function “ResizeLongestSide.”[26]

Zero-shot Performance

The original SAM ViT-base model (model type “vit_b”, checkpoint “sam_vit_b_01ec64.pth ”) was used to evaluate zero-shot performance on the datasets. The larger versions of SAM (ViT Large and ViT Huge) were not used as they did not offer significant performance improvement despite higher computational demands[26]. Bounding box coordinates of each left ventricle segmentation tracing were generated from the ground truth segmentations and used as the prompt for SAM[22].

Model Fine-tuning

The same ViT-base model was used for fine-tuning with a procedure (MedSAM fine-tuning) described by Ma et al (https://github.com/bowang-lab/MedSAM)[24]. We used the training set cases (n=7,460) of the

EchoNet-Dynamic as our customized dataset without further data augmentation. The same bounding box was used as the prompt, as described above. We used a customized loss function, which is the unweighted sum of Dice loss and cross-entropy loss[24,27]. Adam optimizer[28] was used (weight decay= 0), with an initial learning rate of 2e-5 (gradually decreased to 3e-6 over 27 epochs). The batch size was 8. The model was fine-tuned on a node on the Stanford AI lab cluster with a 24 GB NVIDIA RTX TiTAN GPU.

Validation and Generalization

The fine-tuned SAM was tested on the test set of the EchoNet-Dynamic dataset. To test the generalization capacity of SAM, we used external validation samples from the CAMUS dataset (both A2C and A4C) [25] and a Mayo Clinic dataset including the A4C view of cases of TTE and POCUS devices.

Statistical Model Performance Evaluation

The model segmentation performance was directly evaluated by Intersection over Union (IoU) and the Dice similarity coefficient (DSC) against human ground truth labels[29]. Two-tailed, paired t-tests were conducted to assess the statistical significance of the differences between models (zero-shot vs. fine-tuned SAM, and EchoNet vs. fine-tuned SAM), with $p<0.05$ as significant. The summation of disks method was used to calculate the LV ejection fraction (LVEF) from the end-diastolic and end-systolic segmentation masks. LVEF was finally calculated with the following formula.

$$LVEF(\%) = \frac{EDV - ESV}{EDV} \times 100\%$$

While the model was not trained for LVEF prediction, the model-derived LVEF measurements were compared to the ground truth-derived LVEF by R-square and mean absolute error (MAE).

Results

Patient Characteristics

The EchoNet-Dynamic patient characteristics have been described in detail [3]. The mean left ventricular ejection fraction of the EchoNet-dynamic dataset was 55.8±12.4%, 55.8±12.3%, and 55.5±12.2%, for the training, validation, and test set, respectively[3]. In the Mayo Clinic data set (n=99), the mean age was 47.5 ± 17.8 years, 58 (58.6%) were male, and coronary artery disease, hypertension, and diabetes were found in 20 (20.2%), 42 (42.4%), and 15 (15.2%) of patients, respectively. The dataset contains an apical four-chamber view of 58 TTE cases and 41 POCUS cases, the LVEF was 61.7 ± 7.5% for TTE and 63.2 ± 11.9% for POCUS cases. The CAMUS cohort had a mean age of 65.1 ± 14.4 years, 66% male, and a mean LVEF of 44.4 ± 11.9%.

SAM's zero-shot performance on echocardiography and POCUS

Overall, the zero-shot performance on the EchoNet-dynamic test set had a mean DSC of 0.863 ± 0.053. In terms of individual cardiac phase performance, end-diastolic frames were better than end-systolic frames (mean DSC 0.878 ± 0.040 vs. 0.849 ± 0.060). On the Mayo Clinic dataset, the mean DSC was 0.882 ± 0.036 and 0.861 ± 0.043 for TTE and POCUS, respectively. On the CAMUS dataset, we observed mean DSC of 0.866 ± 0.039 and 0.852 ± 0.048 on A4C and A2C views, respectively (**Table 1** and **Supplemental Table 2**). When compared to the ground truth LVEF, the calculated LVEF had an MAE of

11.67%, 6.28%, and 6.38%, on EchoNet, Mayo-TTE, and Mayo-POCUS data, respectively (**Supplemental Table 3**).

Table 1. Comparison of zero-shot SAM, fine-tuned SAM and EchoNet model

Dataset	Phase	DSC					IoU				
		SAM (zero-shot)	SAM (fine-tuned)	EchoNet	p-value*	p-value**	SAM (zero-shot)	SAM (fine-tuned)	EchoNet	p-value*	p-value**
EchoNet-test	Overall	0.863 ± 0.053	0.911 ± 0.045	0.915 ± 0.047	<0.0001	0.0012	0.763 ± 0.077	0.840 ± 0.071	0.847 ± 0.072	<0.0001	0.0003
	ED	0.878 ± 0.040	0.929 ± 0.030	0.903 ± 0.052	<0.0001	<0.0001	0.784 ± 0.062	0.868 ± 0.050	0.826 ± 0.078	<0.0001	<0.0001
	ES	0.849 ± 0.060	0.894 ± 0.050	0.928 ± 0.038	<0.0001	<0.0001	0.742 ± 0.084	0.812 ± 0.077	0.868 ± 0.059	<0.0001	<0.0001
Mayo- TTE	Overall	0.882 ± 0.036	0.902 ± 0.032	0.893 ± 0.090	<0.0001	0.3167	0.790 ± 0.056	0.822 ± 0.051	0.814 ± 0.093	<0.0001	0.3720
	ED	0.889 ± 0.037	0.916 ± 0.024	0.916 ± 0.031	<0.0001	0.9409	0.802 ± 0.058	0.846 ± 0.040	0.846 ± 0.051	<0.0001	0.9991
	ES	0.875 ± 0.033	0.887 ± 0.032	0.870 ± 0.119	0.0039	0.3055	0.779 ± 0.052	0.799 ± 0.050	0.782 ± 0.113	0.0039	0.3114
Mayo-POCUS	Overall	0.861 ± 0.043	0.857 ± 0.047	0.667 ± 0.279	0.4469	<0.0001	0.758 ± 0.066	0.753 ± 0.070	0.554 ± 0.265	0.4695	<0.0001
	ED	0.876 ± 0.032	0.878 ± 0.036	0.717 ± 0.255	0.7355	0.0002	0.781 ± 0.051	0.785 ± 0.056	0.607 ± 0.253	0.7097	<0.0001
	ES	0.846 ± 0.048	0.836 ± 0.047	0.617 ± 0.295	0.2080	<0.0001	0.735 ± 0.072	0.720 ± 0.069	0.501 ± 0.270	0.1994	<0.0001
CAMUS-A2C	Overall	0.852 ± 0.048	0.891 ± 0.040	0.752 ± 0.196	<0.0001	<0.0001	0.745 ± 0.069	0.805 ± 0.062	0.633 ± 0.196	<0.0001	<0.0001
	ED	0.860 ± 0.042	0.897 ± 0.037	0.754 ± 0.196	<0.0001	<0.0001	0.756 ± 0.062	0.815 ± 0.059	0.635 ± 0.197	<0.0001	<0.0001
	ES	0.845 ± 0.052	0.885 ± 0.041	0.751 ± 0.196	<0.0001	<0.0001	0.734 ± 0.073	0.795 ± 0.064	0.632 ± 0.196	<0.0001	<0.0001
CAMUS-A4C	Overall	0.866 ± 0.039	0.897 ± 0.036	0.850 ± 0.097	<0.0001	<0.0001	0.766 ± 0.059	0.815 ± 0.058	0.749 ± 0.117	<0.0001	<0.0001
	ED	0.873 ± 0.037	0.904 ± 0.033	0.850 ± 0.098	<0.0001	<0.0001	0.776 ± 0.056	0.827 ± 0.054	0.749 ± 0.119	<0.0001	<0.0001
	ES	0.860 ± 0.041	0.889 ± 0.038	0.850 ± 0.096	<0.0001	<0.0001	0.756 ± 0.061	0.803 ± 0.060	0.749 ± 0.115	<0.0001	<0.0001

*Zero-shot vs. fine-tuned SAM. **Fine-tuned SAM vs. EchoNet model. A2C: apical 2 chamber view, A4C: apical 4 chamber view, CAMUS: Cardiac Acquisitions for Multi-structure Ultrasound Segmentation, DSC: Dice Similarity Score, ED: end-diastolic, ES: end-systolic, IoU: Intersection over Union, SAM: segment anything model, TTE: transthoracic echocardiography, POCUS: point-of-care ultrasound. Data expressed as mean± standard deviation.

SAM's fine-tuned performance on echocardiography and POCUS

Fine-tuning generally improved the performance of SAM, with a mean DSC of 0.911 ± 0.045 on the EchoNet-dynamic test set (**Table 1** and **Supplemental Table 2**). Similar improvement was also observed in Mayo TTE data, with an overall mean DSC of 0.902 ± 0.032. In contrast, no significant improvement was observed on the POCUS data (DSC 0.857 ± 0.047), while the performance was numerically improved when compared with the ground truth of the second observer (DSC 0.876 ± 0.038), as summarized in **Table 2.** The EchoNet model had a significant performance drop, especially on Mayo-POCUS and CAMUS A2C datasets (**Table 1, Figure 1**). When compared to the ground truth LVEF, the calculated LVEF had an MAE of 7.52%, 5.47%, and 6.70%, on EchoNet, Mayo-TTE, and Mayo-POCUS data, respectively (**Supplemental Table 2**).

Table 2. Zero-shot vs. Fine-tuned SAM performance on TTE and POCUS (against the second observer).

	TTE (n=33)			**POCUS** (n=19)		
	Zero-shot	**Fine-tuned**	**p-value**	**Zero-shot**	**Fine-tuned**	**p-value**
Mean IoU (overall)	0.776 ± 0.063	0.828 ± 0.061	<0.0001	0.755 ± 0.067	0.781 ± 0.059	0.0551
Mean DSC (overall)	0.873 ± 0.040	0.905 ± 0.038	<0.0001	0.859 ± 0.046	0.876 ± 0.038	0.0591
Mean IoU (ED)	0.781 ± 0.057	0.864 ± 0.033	<0.0001	0.773 ± 0.051	0.799 ± 0.052	0.1964
Mean DSC (ED)	0.876 ± 0.037	0.927 ± 0.019	<0.0001	0.871 ± 0.033	0.887 ± 0.032	0.2037
Mean IoU (ES)	0.771 ± 0.068	0.793 ± 0.062	0.1065	0.738 ± 0.078	0.763 ± 0.061	0.1721
Mean DSC (ES)	0.869 ± 0.043	0.883 ± 0.039	0.1044	0.847 ± 0.055	0.864 ± 0.041	0.1794

IoU: intersection over union, DSC: Dice similarity score, ED: end-diastolic, ES: end-systolic. Data expressed as mean± standard deviation.

Discussion

The major contributions of this work include 1) demonstrating the good zero-shot performance of SAM on echocardiography images on the EchoNet-Dynamic, Mayo Clinic, and CAMUS datasets which does not require any training data, and 2) demonstrating the generalization capability of fine-tuned SAM with excellent performance on domain-specific datasets across different institutions and ultrasound modalities. To the best of our knowledge, this is the first work that specifically evaluated the performance of SAM on real-world TTE and POCUS images. Foundation models like SAM are data-efficient options that have the potential to facilitate the development of clinical AI solutions in cardiovascular imaging.

SAM's Zero-shot and fine-tuned performance on echocardiography/POCUS

Echocardiography, like other ultrasound modalities, is generally considered an imaging modality with more challenges due to its operator dependency and low signal-to-noise ratio[3,30–32]. Additionally, objects could often have weak border linings or be obstructed by artifacts on ultrasound/echocardiography images, which posed specific challenges for echocardiography segmentation tasks[3,23]. Compared to other non-cardiac ultrasound images[23], SAM seems to have a relatively better performance on echocardiography with DSC above 0.85 across datasets.

The zero-shot performance of SAM on echocardiography images was good (DSC 0.863 ± 0.053) and was close to the performance of the EchoNet model[3]. On the EchoNet-dynamic dataset, both zero-shot and fine-tuned SAM had a slightly better performance on end-diastolic frames, which was opposite to the original EchoNet model. This could be a result of the fact that end-diastolic frames usually have better visualized left ventricular endocardial borders, and SAM does not depend on the intermediate frames as the video-based EchoNet-Dynamic model[3]. While the frame-based approach does not consider consecutive inter-frame changes and spatio-temporal features of echocardiogram, fine-tuned SAM achieved superior frame-level segmentation performance on non-EchoNet datasets compared to the original video-based EchoNet model[3]. Furthermore, there is no video-based foundation model that handles segmentation tasks like SAM. Importantly, there were cases with suboptimal image quality and

imperfect human labels (**Figure 2**) in the EchoNet-dynamic data set[3,33], which can limit the model performance.

The fine-tuned SAM model demonstrated strong generalization capability, with about a 1% and 5% drop in performance on unseen TTE and POCUS data (**Table 1** and **Figure 1**), respectively. The comparable performance on CAMUS A2C and A4C views also indicated SAM's generalization capability, as it was only fine-tuned on A4C images. Interestingly, the fine-tuned model did not demonstrate an overall superior performance on POCUS images compared to zero-shot performance. This is likely due to inter-observer variation as we observed numerical improvement of the model performance on the other observer's tracings (**Table 2**). Furthermore, on qualitative analysis, the fine-tuned model usually predicts a mask that is more consistent with anticipated LV geometry on POCUS (**Figure 3**).

We also observed a discrepancy between the MAE and R-square metrics on the LVEF assessment. It is important to note that SAM is designed as a universal segmentation model instead of the prediction of LVEF. In contrast, predicting LVEF based on segmented masks is performed by a dedicated model in the EchoNet framework. However, SAM had an MAE ranging from 5-7% on different datasets, which was within the range of typical inter-observer variation, which can be up to 13.9%[10–12].

Integrating SAM and Future Foundation Models into Clinical Research and Practice

One of the major limitations in building machine learning models for healthcare is the collection of high-quality training datasets. In addition, the models' generalizability is frequently questionable[17,34]. Annotators with medical expertise are expensive and limited; while EchoNet is currently the largest public echo segmentation dataset[3], its size is still far from the web-scale data used to train SAM[26]. Having foundation models like SAM can facilitate the development process of image-based AI applications for segmentation tasks[21].

We observed a significant drop in EchoNet's performance on the POCUS dataset (**Table 1**), with completely failed segmentation in 4 (9.8%) of cases. This suggests a limitation of the generalization capacity of conventional neural networks across different modalities. In contrast, the comparable performance of SAM on TTE and POCUS images demonstrated the advantage of leveraging the generalization capabilities of foundation models [21] in building AI solutions for the rapidly growing use of POCUS in cardiac imaging[35]. While evaluation on a larger POCUS is required to better evaluate inter-observer variations, we demonstrated a strategy to fine-tune foundation models using readily available and relatively high-quality TTE results knowing that POCUS images usually come with larger variations in operator skill levels, image quality, and scanning modalities[35–38]. We observed that fine-tuned SAM, similar to MedSAM[24], can accurately segment cases with weak or missing boundaries, which is more common in POCUS images. This can be especially useful in assisting the estimation of cardiac function for POCUS users, who are often not trained for cardiac ultrasound segmentation. SAM can also be used to facilitate the creation of high-quality clinical echocardiography segmentation datasets by reducing the workload of annotation[23,24,26].

We also foresee the potential of incorporating SAM (either zero-shot or fine-tuned) into clinical workflow in echocardiography labs[23]. Compared to the fully automatic models[3], an advantage of having an interactive, human-in-the-loop approach with SAM is that the segmentations can be directly prompted or

modified to the level of interpreters' satisfaction[23,26]. While future studies are needed to assess its real impact on clinical practice, the interactive functionality of SAM could be especially helpful in challenging cases or when fully automatic models fail to accurately predict segmentations[23].

Finally, while the EchoNet-dynamic dataset does not include detailed segmentation other than LV endocardium, a potential future step is to include an object detection model that can detect cardiac chambers, and provide corresponding bounding boxes as prompts to SAM to enable a fully automized AI-segmentation framework.

Conclusion

SAM has a good zero-shot performance on complex echocardiography images from the EchoNet and Mayo Clinic datasets, and its frame-level performance was superior to the original EchoNet model after fine-tuning. The generalization capability of foundation models like SAM can overcome the difficulty of obtaining large-scale, high-quality training data and facilitate the development of AI applications in echocardiography and POCUS.

Limitations

This study is limited by its retrospective nature and could be subject to selection bias. Since SAM is an image-based model, its performance was evaluated on pre-selected end-diastolic and end-systolic frames rather than the beat-to-beat assessment of the video as proposed by the EchoNet-Dynamic model. However, we still demonstrated superior frame-level performance with fine-tuned SAM. While adapters are another potential direction to use SAM for echocardiography segmentation[39], we did not specifically explore this approach in the current paper. How can SAM be integrated into the current echocardiography lab workflow and its real-world effects will need to be validated in a prospective setting.

Ethical review and approval: EchoNet-Dynamic dataset contains a publicly available, de-identified dataset. The use of the Mayo Clinic dataset was approved by the institutional review board (protocol#22-010944).

References

1. Antoine C, Benfari G, Michelena HI, et al. Clinical Outcome of Degenerative Mitral Regurgitation. *Circulation*. 2018;138(13):1317-1326. doi:10.1161/circulationaha.117.033173
2. Matulevicius SA, Rohatgi A, Das SR, Price AL, deLuna A, Reimold SC. Appropriate Use and Clinical Impact of Transthoracic Echocardiography. *JAMA Intern Med*. 2013;173(17):1600-1607. doi:10.1001/jamainternmed.2013.8972
3. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*. 2020;580(7802):252-256. doi:10.1038/s41586-020-2145-8
4. Lang RM, Badano LP, Mor-Avi V, et al. Recommendations for Cardiac Chamber Quantification by Echocardiography in Adults: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *J Am Soc Echocardiogr*. 2015;28(1):1-39.e14. doi:10.1016/j.echo.2014.10.003
5. Liu J, Banchs J, Mousavi N, et al. Contemporary Role of Echocardiography for Clinical Decision Making in Patients During and After Cancer Therapy. *JACC Cardiovasc Imaging*. 2018;11(8):1122-1131. doi:10.1016/j.jcmg.2018.03.025
6. Fonseca R, Jose K, Marwick TH. Understanding decision-making in cardiac imaging: determinants of appropriate use. *European Hear J - Cardiovasc Imaging*. 2017;19(3):262-268. doi:10.1093/ehjci/jex257
7. Otto CM, Nishimura RA, Bonow RO, et al. 2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *J Am Coll Cardiol*. 2020;(J Am Coll Cardiol 63 2014). doi:10.1016/j.jacc.2020.11.018
8. Tam JW, Nichol J, MacDiarmid AL, Lazarow N, Wolfe K. What Is the Real Clinical Utility of Echocardiography? A Prospective Observational Study. J Am Soc Echocardiogr. 1999;12(9):689-697. doi:10.1016/s0894-7317(99)70018-0
9. Pellikka PA, She L, Holly TA, et al. Variability in Ejection Fraction Measured By Echocardiography, Gated Single-Photon Emission Computed Tomography, and Cardiac Magnetic Resonance in Patients With Coronary Artery Disease and Left Ventricular Dysfunction. *JAMA Netw Open*. 2018;1(4):e181456-e181456. doi:10.1001/jamanetworkopen.2018.1456
10. Malm S, Frigstad S, Sagberg E, Larsson H, Skjaerpe T. Accurate and reproducible measurement of left ventricular volume and ejection fraction by contrast echocardiography A comparison with magnetic resonance imaging. *J Am Coll Cardiol*. 2004;44(5):1030-1035. doi:10.1016/j.jacc.2004.05.068
11. Cole GD, Dhutia NM, Shun-Shin MJ, 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*. 2015;31(7):1303-1314. doi:10.1007/s10554-015-0659-1
12. Koh AS, Tay WT, Teng THK, et al. A comprehensive population‐based characterization of heart failure with mid‐range ejection fraction. *Eur J Heart Fail*. 2017;19(12):1624-1634. doi:10.1002/ejhf.945
13. Gilbert A, Marciniak M, Rodero C, Lamata P, Samset E, Mcleod K. Generating Synthetic Labeled Data From Existing Anatomical Models: An Example With Echocardiography Segmentation. *IEEE T Med Imaging*. 2021;40(10):2783-2794. doi:10.1109/tmi.2021.3051806

14. Salte IM, Østvik A, Smistad E, et al. Artificial Intelligence for Automatic Measurement of Left Ventricular Strain in Echocardiography. JACC Cardiovasc Imaging. 2021;14(10):1918-1928. doi:10.1016/j.jcmg.2021.04.018
15. Huang H, Ge Z, Wang H, et al. Segmentation of Echocardiography Based on Deep Learning Model. *Electronics*. 2022;11(11):1714. doi:10.3390/electronics11111714
16. Amer A, Ye X, Janan F. ResDUnet: A Deep Learning-Based Left Ventricle Segmentation Method for Echocardiography. *IEEE Access*. 2021;9:159755-159763. doi:10.1109/access.2021.3122256
17. Yu Y, Wang C, Fu Q, et al. Techniques and Challenges of Image Segmentation: A Review. *Electronics*. 2023;12(5):1199. doi:10.3390/electronics12051199
18. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *Arxiv*. Published online 2020. doi:10.48550/arxiv.2010.11929
19. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. *Arxiv*. Published online 2017. doi:10.48550/arxiv.1706.03762
20. Liu Y, Han T, Ma S, et al. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. *Arxiv*. Published online 2023.
21. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. Arxiv. Published online 2021. doi:10.48550/arxiv.2108.07258
22. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. *Arxiv*. Published online 2023.
23. Mazurowski MA, Dong H, Gu H, Yang J, Konz N, Zhang Y. Segment Anything Model for Medical Image Analysis: an Experimental Study. *Arxiv*. Published online 2023.
24. Ma J, Wang B. Segment Anything in Medical Images. *Arxiv*. Published online 2023. doi:10.48550/arxiv.2304.12306
25. Leclerc S, Smistad E, Pedrosa J, et al. Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. *IEEE Trans Méd Imaging*. 2019;38(9):2198-2210. doi:10.1109/tmi.2019.2900516
26. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. *Arxiv*. Published online 2023.
27. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z
28. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *Arxiv*. Published online 2014. doi:10.48550/arxiv.1412.6980
29. Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: Recommendations for image analysis validation. arXiv. Published online 2022. doi:10.48550/arxiv.2206.01653
30. Mintz GS, Kotler MN. Clinical Value and Limitations of Echocardiography: Its Use in the Study of Patients With Infectious Endocarditis. *Arch Intern Med*. 1980;140(8):1022-1027. doi:10.1001/archinte.1980.00330190034014
31. Mondillo S, Maccherini M, Galderisi M. Usefulness and limitations of transthoracic echocardiography in heart transplantation recipients. *Cardiovasc Ultrasound*. 2008;6(1):2-2. doi:10.1186/1476-7120-6-2
32. Abdulla AM, Frank MJ, Canedo MI, Stefadouros MA. Limitations of echocardiography in the assessment of left ventricular size and function in aortic regurgitation. *Circulation*. 2018;61(1):148-155. doi:10.1161/01.cir.61.1.148
33. Tromp J, Seekings PJ, Hung CL, et al. Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. *Lancet Digital Heal*. 2022;4(1):e46-e54. doi:10.1016/s2589-7500(21)00235-1

34. Zhang D, Lin Y, Chen H, et al. Deep Learning for Medical Image Segmentation: Tricks, Challenges and Future Directions. *Arxiv*. Published online 2022. doi:10.48550/arxiv.2209.10307
35. Huang GS, Alviar CL, Wiley BM, Kwon Y. The Era of Point-of-Care Ultrasound Has Arrived: Are Cardiologists Ready? *Am J Cardiol*. 2020;132:173-175. doi:10.1016/j.amjcard.2020.06.062
36. Kalagara H, Coker B, Gerstein NS, et al. Point-of-Care Ultrasound (POCUS) for the Cardiothoracic Anesthesiologist. *J Cardiothor Vasc An*. 2022;36(4):1132-1147. doi:10.1053/j.jvca.2021.01.018
37. Palmero SL, Zúñiga MAL, Martínez VR, et al. Point-of-Care Ultrasound (POCUS) as an Extension of the Physical Examination in Patients with Bacteremia or Candidemia. *J Clin Medicine*. 2022;11(13):3636. doi:10.3390/jcm11133636
38. Kline J, Golinski M, Selai B, Horsch J, Hornbaker K. The effectiveness of a blended POCUS curriculum on achieving basic focused bedside transthoracic echocardiography (TTE) proficiency. A formalized pilot study. *Cardiovasc Ultrasoun*. 2021;19(1):39. doi:10.1186/s12947-021-00268-9
39. Wu J, Fu R, Fang H, et al. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *Arxiv*. Published online 2023.

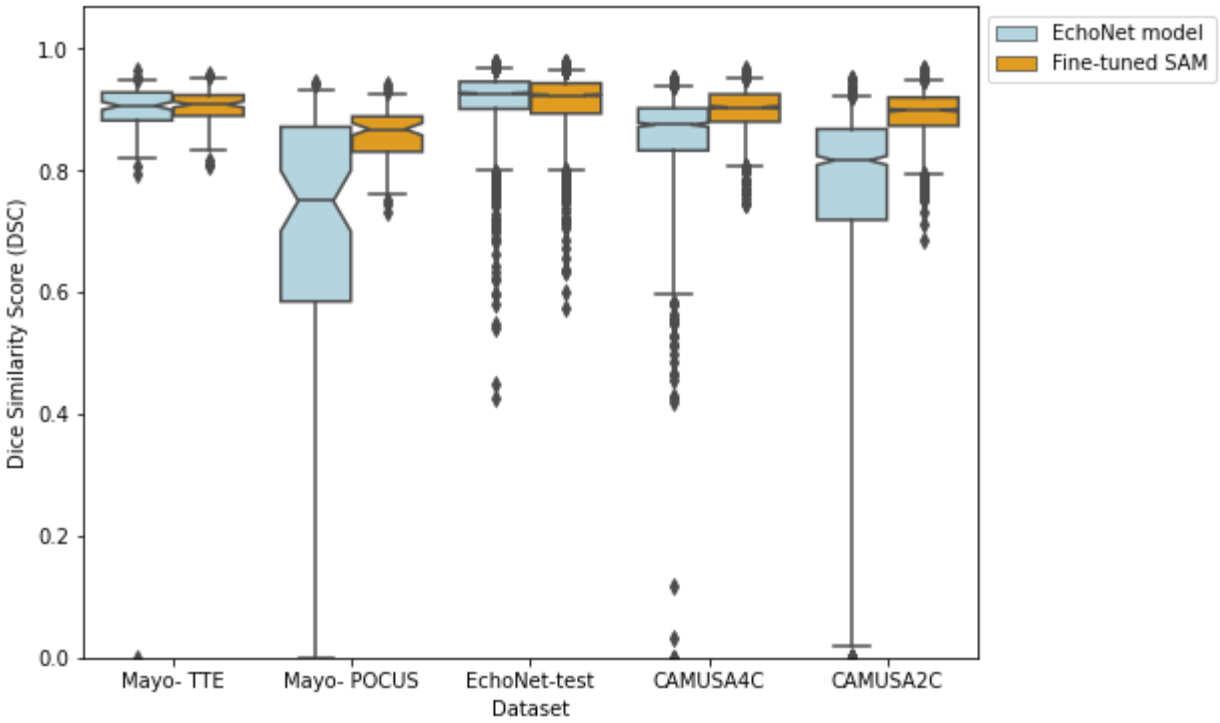

Figure 1. Comparison of the overall Dice Similarity Scores (DSC) between EchoNet and fine-tuned SAM models. This figure illustrates the distribution of overall DSC scores for EchoNet (in light blue) and Fine-tuned SAM (in orange), across multiple datasets. The box plots depict the median, quartiles, and 95% confidence intervals of the overall DSC scores. Fine-tuned SAM demonstrated consistent performance across different datasets and was significantly superior to the EchoNet model on the Mayo-TTE, Mayo-POCUS, and CAMUS datasets (all $p<0.0001$).

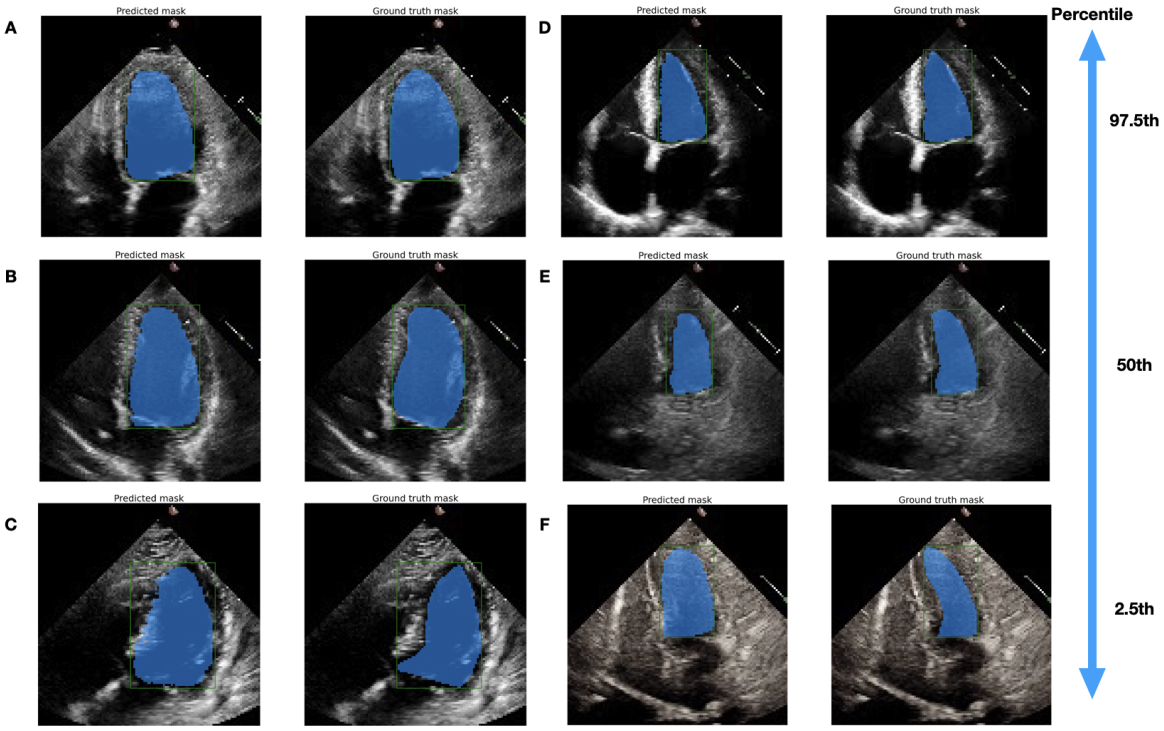

Figure 2. Qualitative performance of fine-tuned SAM on representative cases against ground truth on the EchoNet-dynamic test dataset. From top to bottom: 97.5th to 2.5th percentile of DSC. Panels A, B, and C are end-diastolic frames, and Panels D, E, and F are end-systolic frames. We observed that many of the poor-performance cases had suboptimal image qualities, such as weak LV borders or off-axis views (Panels C and F), suggesting the importance of good input image quality on model performance. Additionally, end-diastolic frames usually have a better delineation of borders than end-systolic frames, which is consistent with the model performance (end-diastolic slightly better than end-systolic).

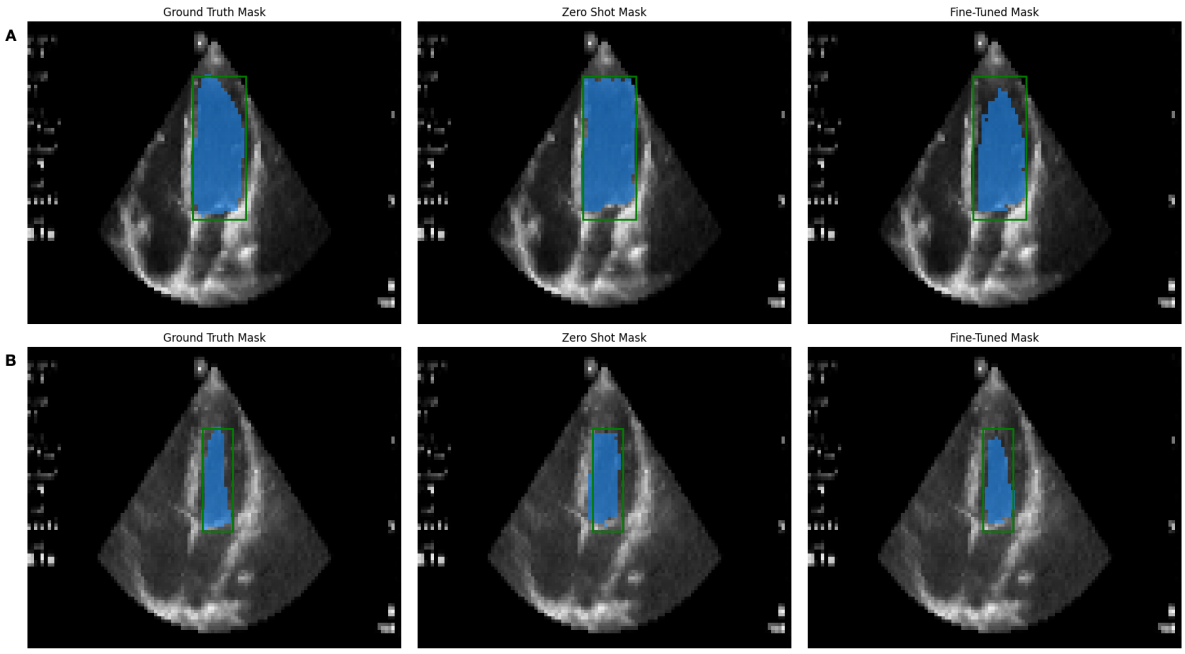

Figure 3. Zero-shot and fine-tuned SAM performance on a representative POCUS case. **Panel A.** end-diastolic frame, **Panel B**. end-systolic frame. From left to right are the ground truth, zero-shot, and fine-tuned mask, with an overlay of bounding boxes (green-colored) and mask (blue-colored), on the original POCUS image. Fine-tuned masks were more consistent with anticipated left ventricular geometry on visualization. Note that POCUS images generally had worse quality compared to transthoracic echocardiography images.

Supplemental Table 1. Excluded EchoNet-dynamic dataset cases due to missing ground truth

FileName	Split
0X234005774F4CB5CD.avi	TRAIN
0X2DC68261CBCC04AE.avi	TRAIN
0X35291BE9AB90FB89.avi	TRAIN
0X6C435C1B417FDE8A.avi	TRAIN
0X5515B0BD077BE68A.avi	TRAIN
0X5DD5283AC43CCDD1.avi	TEST

Supplemental Table 2. Zero-shot and Fine-tuned SAM performance on EchoNet-dynamic Train and Validation set

	Zero-shot		**Fine-tuned**	
	EchoNet-Dynamic-Train	**EchoNet-Dynamic-Validation**	**EchoNet-Dynamic-Train**	**EchoNet-Dynamic-Validation**
Mean IoU (overall)	0.761 ± 0.080	0.762 ± 0.078	0.842 ± 0.070	0.841 ± 0.069
Mean DSC (overall)	0.862 ± 0.055	0.862 ± 0.054	0.913 ± 0.044	0.912 ± 0.043
Mean IoU (ED)	0.781 ± 0.068	0.783 ± 0.065	0.868 ± 0.052	0.869 ± 0.050
Mean DSC (ED)	0.875 ± 0.046	0.877 ± 0.043	0.928 ± 0.032	0.929 ± 0.030
Mean IoU (ES)	0.740 ± 0.085	0.740 ± 0.084	0.816 ± 0.076	0.813 ± 0.074
Mean DSC (ES)	0.848 ± 0.060	0.848 ± 0.059	0.897 ± 0.049	0.895 ± 0.047
IoU: intersection over union, DSC: Dice similarity score, ED: end-diastolic, ES: end-systolic. Data expressed as mean± standard deviation.				

Supplemental Table 3. LVEF prediction task evaluated by R-square and MAE of zero-shot vs. fine-tuned SAM

	EchoNet-dynamic test		**CAMUS-A2C**		**CAMUS-A4C**		**Mayo-TTE**		**Mayo-POCUS**	
	Zero-shot	**Fine-tuned**	**Zero-shot**	**Fine-tuned**	**Zero-shot**	**Fine-tuned**	**Zero-shot**	**Fine-tuned**	**Zero-shot**	**Fine-tuned**
R-squared	0.141	0.161	0.445	0.761	0.549	0.709	0.374	0.517	0.638	0.718
MAE (%)	11.7	7.52	9.75	6.11	9.67	7.31	6.28	5.47	6.38	6.70
MAE: mean absolute error.										