Title: Foundation versus Domain-Specific Models for Left Ventricular Segmentation on Cardiac Ultrasound

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

[*]Li Fei-Fei and Ehsan Adeli are co-corresponding authors of this manuscript.

Affiliations:

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

Corresponding Author:
Ehsan Adeli, Ph.D., Assistant Professor
Department of Psychiatry and Behavioral Sciences
Email: eadeli@stanford.edu
Address: MC#5590, 1070 Arastradero Rd., Palo Alto, CA 94304

Abstract

The Segment Anything Model (SAM) was fine-tuned on the EchoNet-Dynamic dataset and evaluated on external transthoracic echocardiography (TTE) and Point-of-Care Ultrasound (POCUS) datasets from CAMUS (University Hospital of St Etienne) and Mayo Clinic (99 patients: 58 TTE, 41 POCUS). Fine-tuned SAM was superior or comparable to MedSAM. The fine-tuned SAM also outperformed EchoNet and U-Net models, demonstrating strong generalization, especially on apical 2-chamber (A2C) images (fine-tuned SAM vs. EchoNet: 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, $p<0.0001$). Additionally, SAM-enhanced workflow reduced annotation time by 50% (11.6 ± 4.5 sec vs. 5.7 ± 1.7 sec, $p<0.0001$) while maintaining segmentation quality. We demonstrated an effective strategy for fine-tuning a vision foundation model for enhancing clinical workflow efficiency and supporting human-AI collaboration.

Introduction

Echocardiography provides a 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, especially the left ventricle (LV), 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 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[3,13–18].

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 costly to implement[19]. In recent years, transformers, a type of neural network architecture with the self-attention mechanism that enables the model to efficiently capture complex dependencies and relationships, have revolutionized the field from natural language processing to computer vision[20–22]. Vision transformers (ViT) [20] 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 applications[23]. 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, i.e., segments user-specified objects at different data resources without needing any training data[24].

However, while the zero-shot performance of SAM on natural image datasets has been promising[24], 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[25]. Recently, MedSAM was introduced as a universal tool for medical image segmentation. However, ultrasound or echocardiography was less represented in the training set despite > 1 million medical images being used[26]. This prompted a core research question: How can we effectively develop a segmentation model for a specific medical image segmentation task? Is it necessary to build a universal medical image segmentation model, or can a generic foundation model be fine-tuned with a relatively small, task-specific dataset to achieve comparable performance? In this context, we aim to study a novel strategy of using representative echocardiography examples for the fine-tuning of a foundation segmentation model (SAM) and compare its effectiveness with MedSAM and a state-of-the-art segmentation model trained with a domain-specific dataset (EchoNet)[3].

We hypothesized that fine-tuning the foundation segmentation model with the above strategy can achieve similar or superior performance on LV segmentation compared to MedSAM or EchoNet. We further designed a human annotation study on a subset of images to investigate whether the model can enhance the efficiency of LV segmentation. 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 on external evaluation sets.

Results

SAM's zero-shot performance on echocardiography and POCUS

We first tested the zero-shot performance of SAM. 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) (**Table 1**). **Supplementary Table 1** summarizes the zero-shot performance on the EchoNet-dynamic training and validation sets. 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 a mean DSC of 0.866 ± 0.039 and 0.852 ± 0.048 on A4C and A2C views, respectively (**Table 1**). 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 (**Supplementary Table 2**).

Table 1. Comparison of Model Performance Based on Dice Similarity Coefficient (DSC) and Intersection over Union (IoU)

Dataset	Phase	DSC					IoU				
		SAM_{112} (fine-tuned)	SAM (zero-shot)	MedSAM	EchoNet	U-Net	SAM_{112} (fine-tuned)	SAM (zero-shot)	MedSAM	EchoNet	U-Net
EchoNet-test	Overall	0.911 ± 0.045$^{*\dagger\emptyset+}$	0.863 ± 0.053	0.872 ± 0.056	0.915 ± 0.047	0.640 ± 0.151	0.840 ± 0.071$^{*\dagger\emptyset+}$	0.763 ± 0.077	0.777 ± 0.083	0.847 ± 0.072	0.488 ± 0.157
	ED	0.929 ± 0.030$^{*\dagger\emptyset+}$	0.878 ± 0.040	0.893 ± 0.040	0.903 ± 0.052	0.675 ± 0.134	0.868 ± 0.050$^{*\dagger\emptyset+}$	0.784 ± 0.062	0.809 ± 0.062	0.826 ± 0.078	0.524 ± 0.148
	ES	0.894 ± 0.050$^{*\dagger\emptyset+}$	0.849 ± 0.060	0.850 ± 0.061	0.928 ± 0.038	0.606 ± 0.160	0.812 ± 0.077$^{*\dagger\emptyset+}$	0.742 ± 0.084	0.744 ± 0.088	0.868 ± 0.059	0.453 ± 0.159
Mayo-TTE	Overall	0.902 ± 0.032$^{*\emptyset+}$	0.882 ± 0.036	0.883 ± 0.039	0.893 ± 0.090	0.597 ± 0.169	0.822 ± 0.051$^{*\emptyset+}$	0.790 ± 0.056	0.792 ± 0.061	0.814 ± 0.093	0.444 ± 0.162
	ED	0.916 ± 0.024$^{*\emptyset+}$	0.889 ± 0.037	0.903 ± 0.027	0.916 ± 0.031	0.634 ± 0.162	0.846 ± 0.040$^{*\emptyset+}$	0.802 ± 0.058	0.825 ± 0.044	0.846 ± 0.051	0.482 ± 0.161
	ES	0.887 ± 0.032$^{*\emptyset+}$	0.875 ± 0.033	0.862 ± 0.039	0.870 ± 0.119	0.561 ± 0.169	0.799 ± 0.050$^{*\emptyset+}$	0.779 ± 0.052	0.760 ± 0.059	0.782 ± 0.113	0.408 ± 0.156
Mayo-POCUS	Overall	0.857 ± 0.047$^{\dagger+}$	0.861 ± 0.043	0.846 ± 0.072	0.667 ± 0.279	0.396 ± 0.239	0.753 ± 0.070$^{\dagger+}$	0.758 ± 0.066	0.740 ± 0.102	0.554 ± 0.265	0.275 ± 0.192
	ED	0.878 ± 0.036$^{\dagger+}$	0.876 ± 0.032	0.868 ± 0.064	0.717 ± 0.255	0.463 ± 0.232	0.785 ± 0.056$^{\dagger+}$	0.781 ± 0.051	0.772 ± 0.092	0.607 ± 0.253	0.329 ± 0.194
	ES	0.836 ± 0.047$^{\dagger+}$	0.846 ± 0.048	0.825 ± 0.074	0.617 ± 0.295	0.330 ± 0.229	0.720 ± 0.069$^{\dagger+}$	0.735 ± 0.072	0.708 ± 0.103	0.501 ± 0.270	0.221 ± 0.175
CAMUS-A2C	Overall	0.891 ± 0.040$^{*\dagger\emptyset+}$	0.852 ± 0.048	0.857 ± 0.057	0.752 ± 0.196	0.460 ± 0.154	0.805 ± 0.062$^{*\dagger\emptyset+}$	0.745 ± 0.069	0.754 ± 0.081	0.633 ± 0.196	0.311 ± 0.130
	ED	0.897 ± 0.037$^{*\dagger\emptyset+}$	0.860 ± 0.042	0.877 ± 0.040	0.754 ± 0.196	0.458 ± 0.135	0.815 ± 0.059$^{*\dagger\emptyset+}$	0.756 ± 0.062	0.783 ± 0.060	0.635 ± 0.197	0.307 ± 0.116
	ES	0.885 ± 0.041$^{*\dagger\emptyset+}$	0.845 ± 0.052	0.838 ± 0.065	0.751 ± 0.196	0.461 ± 0.172	0.795 ± 0.064$^{*\dagger\emptyset+}$	0.734 ± 0.073	0.725 ± 0.088	0.632 ± 0.196	0.315 ± 0.144
CAMUS-A4C	Overall	0.897 ± 0.036$^{*\dagger\emptyset+}$	0.866 ± 0.039	0.877 ± 0.044	0.850 ± 0.097	0.418 ± 0.159	0.815 ± 0.058$^{*\dagger\emptyset+}$	0.766 ± 0.059	0.784 ± 0.066	0.749 ± 0.117	0.277 ± 0.128
	ED	0.904 ± 0.033 $^{*\dagger\emptyset+}$	0.873 ± 0.037	0.890 ± 0.035	0.850 ± 0.098	0.428 ± 0.139	0.827 ± 0.054$^{*\dagger\emptyset+}$	0.776 ± 0.056	0.804 ± 0.055	0.749 ± 0.119	0.282 ± 0.114
	ES	0.889 ± 0.038 $^{*\dagger\emptyset+}$	0.860 ± 0.041	0.864 ± 0.048	0.850 ± 0.096	0.409 ± 0.176	0.803 ± 0.060$^{*\dagger\emptyset+}$	0.756 ± 0.061	0.763 ± 0.070	0.749 ± 0.115	0.272 ± 0.140

*p< 0.05, Zero-shot vs. fine-tuned SAM. †p< 0.05, Fine-tuned SAM vs. EchoNet model. $^{\emptyset}$p< 0.05, Fine-tuned SAM vs. MedSAM. $^{+}$p< 0.05, Fine-tuned SAM vs. U-Net. 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.

Table 2. Comparison of Model Performance Based on Hausdorff Distance (HD) and Average Surface Distance (ASD)

Dataset	Phase	HD	ASD

		SAM_{112} (fine-tuned)	SAM (zero-shot)	MedSAM	EchoNet	U-Net	SAM_{112} (fine-tuned)	SAM (zero-shot)	MedSAM	EchoNet	U-Net
Echo Net-test	**Overall**	$4.408 \pm 1.999^{*\dagger\emptyset+}$	7.780 ± 3.155	6.381 ± 2.143	9.196 ± 2.929	13.596 ± 6.131	$0.156 \pm 0.133^{*\dagger\emptyset+}$	0.350 ± 0.211	0.286 ± 0.181	0.753 ± 0.371	1.785 ± 1.429
	ED	$4.425 \pm 2.066^{*\dagger\emptyset+}$	8.320 ± 3.274	6.490 ± 2.272	9.715 ± 2.966	14.060 ± 6.637	$0.123 \pm 0.098^{*\dagger\emptyset+}$	0.338 ± 0.202	0.240 ± 0.150	0.860 ± 0.395	1.668 ± 1.340
	ES	$4.390 \pm 1.931^{*\dagger\emptyset+}$	7.239 ± 2.935	6.271 ± 2.000	8.678 ± 2.799	13.132 ± 5.543	$0.188 \pm 0.154^{*\dagger\emptyset+}$	0.361 ± 0.219	0.331 ± 0.198	0.646 ± 0.309	1.903 ± 1.505
Mayo-TTE	**Overall**	$7.543 \pm 4.780^{*\dagger+}$	8.581 ± 4.495	$4.435 \pm 1.275^{\emptyset}$	23.714 ± 6.585	13.896 ± 5.261	$0.814 \pm 0.930^{*\dagger+}$	0.854 ± 0.928	$0.401 \pm 0.147^{\emptyset}$	5.705 ± 2.984	1.877 ± 1.425
	ED	$7.855 \pm 5.225^{*\dagger+}$	9.127 ± 4.847	$4.920 \pm 1.402^{\emptyset}$	26.539 ± 6.527	13.067 ± 4.669	$0.903 \pm 1.061^{*\dagger+}$	0.927 ± 1.011	$0.412 \pm 0.165^{\emptyset}$	5.337 ± 2.502	1.585 ± 1.214
	ES	$7.241 \pm 4.333^{*\dagger+}$	8.054 ± 4.101	$3.951 \pm 0.914^{\emptyset}$	20.938 ± 5.389	14.697 ± 5.701	$0.728 \pm 0.783^{\dagger+}$	0.785 ± 0.843	$0.390 \pm 0.126^{\emptyset}$	6.066 ± 3.375	2.159 ± 1.562
Mayo-POCUS	**Overall**	$8.804 \pm 6.716^{*\dagger+}$	9.072 ± 6.291	$3.666 \pm 1.041^{\emptyset}$	19.821 ± 5.776	24.452 ± 23.324	$1.202 \pm 1.311^{\dagger+}$	1.172 ± 1.297	$0.436 \pm 0.234^{\emptyset}$	5.421 ± 2.655	4.911 ± 5.545
	ED	$8.273 \pm 6.395^{*\dagger+}$	9.033 ± 5.851	$3.858 \pm 1.035^{\emptyset}$	20.059 ± 5.617	20.855 ± 19.624	$0.972 \pm 1.113^{\dagger+}$	1.065 ± 1.248	$0.362 \pm 0.150^{\emptyset}$	4.889 ± 2.061	3.792 ± 4.414
	ES	$9.335 \pm 7.061^{\dagger+}$	9.111 ± 6.776	$3.475 \pm 1.023^{\emptyset}$	19.595 ± 5.983	28.050 ± 26.268	$1.431 \pm 1.462^{*\dagger+}$	1.279 ± 1.351	$0.511 \pm 0.278^{\emptyset}$	5.927 ± 3.058	6.029 ± 6.341
CAMUS-A2C	**Overall**	$5.624 \pm 2.121^{*\dagger+}$	5.842 ± 2.081	$5.169 \pm 2.354^{\emptyset}$	6.962 ± 3.951	18.709 ± 7.665	$0.237 \pm 0.147^{*\dagger+}$	0.257 ± 0.154	$0.220 \pm 0.214^{\emptyset}$	0.469 ± 0.482	3.277 ± 2.000
	ED	$6.120 \pm 2.198^{*+}$	6.302 ± 2.071	$5.294 \pm 2.365^{\emptyset}$	6.653 ± 3.676	20.610 ± 8.075	$0.236 \pm 0.155^{*\dagger+}$	0.252 ± 0.131	$0.193 \pm 0.155^{\emptyset}$	0.384 ± 0.294	3.412 ± 1.907
	ES	$5.129 \pm 1.920^{*\dagger+}$	5.382 ± 1.990	$5.043 \pm 2.339^{\emptyset}$	7.270 ± 4.190	16.808 ± 6.720	$0.237 \pm 0.139^{*\dagger+}$	0.262 ± 0.173	$0.248 \pm 0.257^{\emptyset}$	0.555 ± 0.603	3.142 ± 2.081
CAMUS-A4C	**Overall**	$5.001 \pm 1.728^{*\dagger+}$	5.601 ± 1.926	$4.826 \pm 2.515^{\emptyset}$	5.763 ± 2.519	19.748 ± 7.868	$0.205 \pm 0.117^{*\dagger+}$	0.230 ± 0.124	$0.176 \pm 0.173^{\emptyset}$	0.370 ± 0.267	3.717 ± 1.938
	ED	$5.367 \pm 1.850^{*\dagger+}$	6.085 ± 1.919	$5.223 \pm 2.895^{\emptyset}$	5.838 ± 2.523	21.822 ± 8.322	$6.120 \pm 2.198^{+}$	6.302 ± 2.071	$5.294 \pm 2.365^{\emptyset}$	6.653 ± 3.676	20.610 ± 8.075
	ES	$4.635 \pm 1.512^{*\dagger+}$	5.118 ± 1.809	$4.428 \pm 1.990^{\emptyset}$	5.687 ± 2.515	17.674 ± 6.787	$5.129 \pm 1.920^{*\dagger+}$	5.382 ± 1.990	$5.043 \pm 2.339^{\emptyset}$	7.270 ± 4.190	16.808 ± 6.720

*$p < 0.05$, Zero-shot vs. fine-tuned SAM. †$p < 0.05$, Fine-tuned SAM vs. EchoNet model. $^{\emptyset}p < 0.05$, Fine-tuned SAM vs. MedSAM. $^{+}p < 0.05$, Fine-tuned SAM vs. U-Net. A2C: apical 2 chamber view, A4C: apical 4 chamber view, CAMUS: Cardiac Acquisitions for Multi-structure Ultrasound Segmentation, ASD: Average Surface Distance, HD: Hausdorff Distance, 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

Compared to zero-shot, fine-tuning generally improved the performance of SAM, with a mean DSC of 0.911 ± 0.045 (SAM_{112}) on the EchoNet-dynamic test set (**Table 1** and **Supplementary Table 1**). 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 3.** The EchoNet model had a significant performance drop, especially on the Mayo-POCUS and CAMUS A2C datasets (**Table 1**). When compared to the ground truth LVEF, the calculated LVEF had an MAE of 7.52%, 5.47%, and 6.70%, on the EchoNet test, Mayo-TTE, and Mayo-POCUS data, respectively (**Supplementary Table 2**).

Qualitatively, poor-performing cases often featured suboptimal images (such as weak LV endocardial borders, off-axis views; **Figure 1**, Panels c, f), highlighting the importance of image quality (**Figure 1**). On a representative POCUS case, fine-tuned SAM_{112} predicted masks that were more consistent with LV geometry (**Figure 2**).

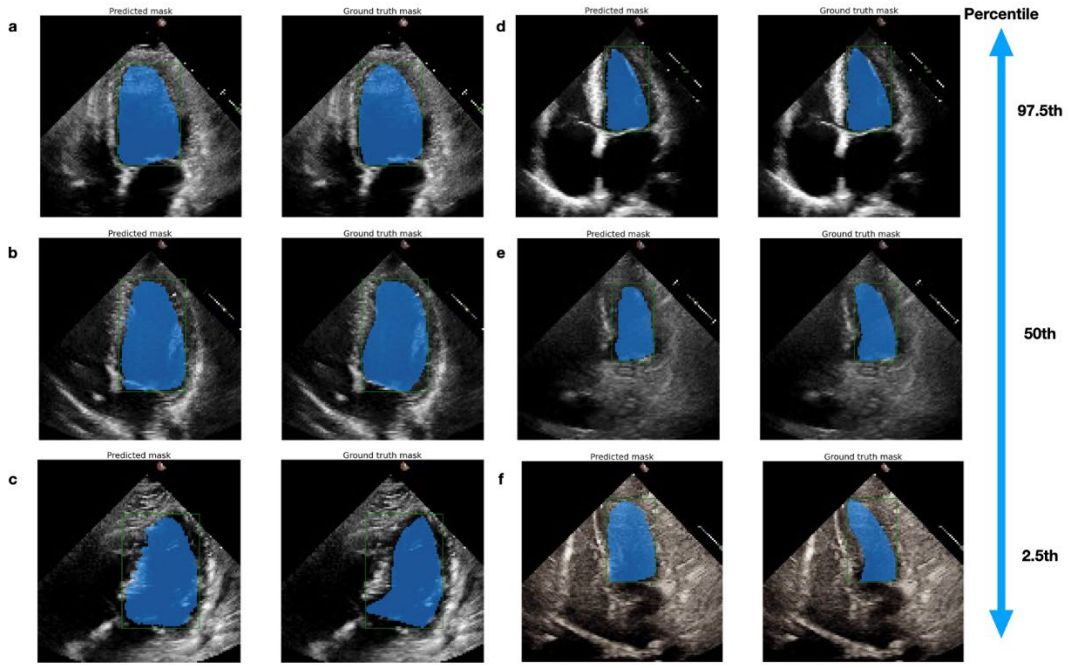

Figure 1. Qualitative performance of fine-tuned SAM on representative cases.

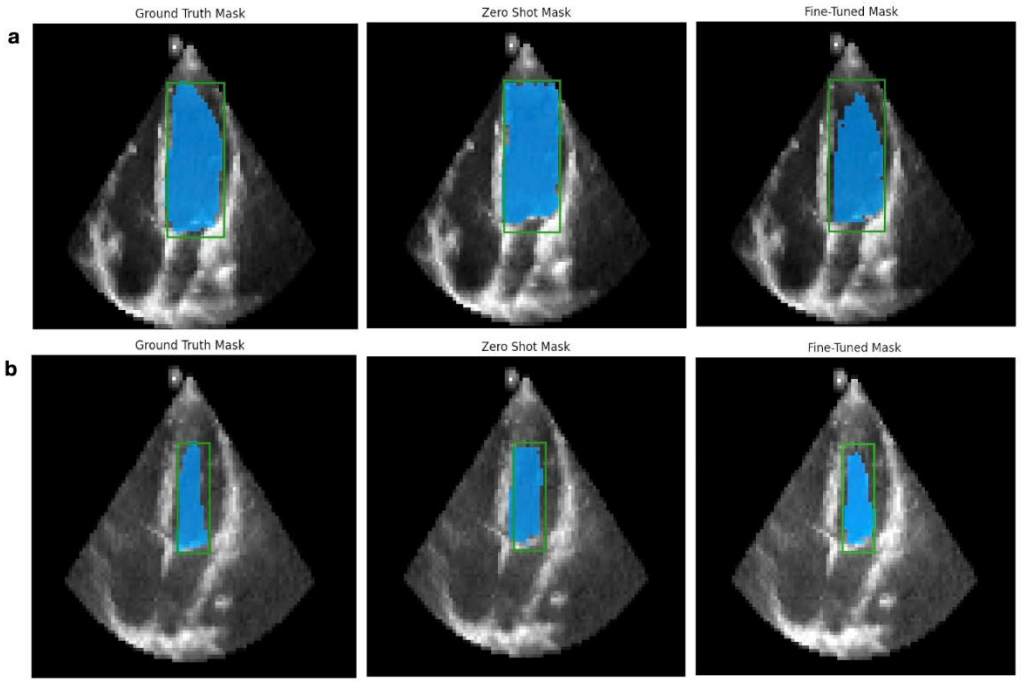

Figure 2. Zero-shot and fine-tuned SAM performance on a representative POCUS case.

Table 3. Zero-shot vs. Fine-tuned SAM_{112} performance on TTE and POCUS (against the second observer).

	TTE (n=33)			**POCUS** (n=19)		
	Zero-shot	**Fine-tuned** SAM_{112}	**p-value**	**Zero-shot**	**Fine-tuned** SAM_{112}	**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.

Comparison of SAM, MedSAM, EchoNet, and U-Net models

Comparisons were made between fine-tuned SAM_{112} (fine-tuned using 112× 112 images), MedSAM, EchoNet, and U-Net. Under the same setting, the inference time for SAM was about 7.3 images/sec, MedSAM was 5.0 images/sec, EchoNet was 153.3 images/sec, and UNet was 26 images/sec. Note that EchoNet takes video input, so the inference time was averaged by the number of frames in each video.

When evaluated by DSC and IoU, EchoNet demonstrated a similar level of performance compared to fine-tuned SAM_{112} on the EchoNet test set (SAM_{112} vs. EchoNet: DSC 0.911 ± 0.045 vs. 0.915 ± 0.047, p<0.0001) and the Mayo TTE dataset (DSC 0.902 ± 0.032 vs. 0.893 ± 0.090, p<0.0001). However, EchoNet significantly underperformed on the Mayo POCUS and CAMUS datasets, with a performance drop ranging from 5-25% (**Table 1**). MedSAM also showed about 2-5% worse performance than fine-tuned SAM_{112}across all datasets (all p<0.0001), except for the CAMUS dataset, as it was part of MedSAM's training data. U-Net had worse performance than fine-tuned SAM_{112} across all datasets (all p< 0.0001) (**Table 1**). When evaluated by HD and ASD, MedSAM had the best performance across most of the datasets, while fine-tuned SAM_{112} had the best performance on the EchoNet test set. This was followed by SAM (zero-shot), EchoNet, and then U-Net (**Supplementary Table 3**).

Further fine-tuned SAM_{1024} (fine-tuned using 1024× 1024 images as MedSAM) essentially achieved the same performance as fine-tuned MedSAM, with a DSC of 0.935 and 0.936, respectively. Both of the fine-tuned models also demonstrated similar performance across different datasets, except for the CAMUS dataset, which is part of MedSAM's training set (**Table 4**). We also observed that after the finetuning, both SAM_{1024} and MedSAM had dropped performance slightly on the Mayo and POCUS datasets, compared to fine-tuned SAM_{112} and base MedSAM.

Table 4. Comparison of fine-tuned SAM and MedSAM

Dataset	**Phase**	**DSC**		**IoU**		**ASD**		**HD**	
		SAM_{1024} (fine-tuned)	**MedSAM (fine-tuned)**	**SAM_{1024} (fine-tuned)**	**MedSAM (fine-tuned)**	**SAM_{1024} (fine-tuned)**	**MedSAM (fine-tuned)**	**SAM_{1024} (fine-tuned)**	**MedSAM (fine-tuned)**
EchoNet-test	**Overall**	0.935 ± 0.031	0.936 ± 0.030*	0.879 ± 0.051	0.881 ± 0.050*	0.089 ± 0.071	0.087 ± 0.061	3.041 ± 1.158	3.016 ± 1.105
	ED	0.946 ± 0.021	0.947 ± 0.020*	0.898 ± 0.036	0.900 ± 0.035*	0.074 ± 0.044	0.071 ± 0.042*	3.084 ± 1.105	3.046 ± 1.095
	ES	0.924 ± 0.035	0.924 ± 0.034	0.861 ± 0.057	0.861 ± 0.055	0.104 ± 0.088	0.103 ± 0.072	2.998 ± 1.206	2.986 ± 1.115
Mayo-TTE	**Overall**	0.899 ± 0.026	0.894 ± 0.029	0.817 ± 0.042	0.810 ± 0.046	0.136 ± 0.044	0.142 ± 0.052	3.232 ± 0.743	3.294 ± 0.763
	ED	0.913 ± 0.016	0.911 ± 0.017	0.839 ± 0.026	0.838 ± 0.028	0.116 ± 0.028	0.117 ± 0.026	3.127 ± 0.715	3.220 ± 0.641
	ES	0.885 ± 0.027	0.877 ± 0.029	0.795 ± 0.044	0.782 ± 0.044	0.155 ± 0.048	0.167 ± 0.059	3.336 ± 0.763	3.369 ± 0.867
Mayo-	**Overall**	0.842 ± 0.056	0.839 ± 0.062	0.731 ± 0.075	0.727 ± 0.086	0.229 ± 0.217	0.240 ± 0.244	3.289 ± 1.424	3.373 ± 1.645
	ED	0.860 ± 0.060	0.862 ± 0.052	0.758 ± 0.078	0.761 ± 0.071	0.223 ± 0.299	0.224 ± 0.321	3.317 ± 1.857	3.375 ± 2.134

POCUS	**ES**	0.824 ± 0.044	0.815 ± 0.063	0.703 ± 0.062	0.693 ± 0.086	0.235 ± 0.073	0.256 ± 0.128	3.260 ± 0.809	3.371 ± 0.962
CAMUS-A2C	**Overall**	0.845 ± 0.063	0.872 ± 0.062*	0.736 ± 0.088	0.777 ± 0.085*	0.392 ± 0.320	0.288 ± 0.309*	6.727 ± 2.983	5.758 ± 2.927*
	ED	0.861 ± 0.050	0.886 ± 0.043*	0.760 ± 0.075	0.798 ± 0.066*	0.364 ± 0.284	0.259 ± 0.228*	7.008 ± 3.036	6.032 ± 3.035*
	ES	0.828 ± 0.070	0.857 ± 0.073*	0.712 ± 0.094	0.756 ± 0.096*	0.420 ± 0.351	0.316 ± 0.371*	6.446 ± 2.905	5.484 ± 2.791*
CAMUS-A4C	**Overall**	0.861 ± 0.051	0.880 ± 0.047*	0.759 ± 0.075	0.790 ± 0.072*	0.323 ± 0.261	0.255 ± 0.239*	6.075 ± 2.812	5.419 ± 2.807*
	ED	0.869 ± 0.049	0.888 ± 0.045*	0.772 ± 0.073	0.801 ± 0.069*	0.336 ± 0.300	0.260 ± 0.271*	6.709 ± 3.184	5.862 ± 3.196*
	ES	0.853 ± 0.051	0.873 ± 0.049*	0.747 ± 0.075	0.778 ± 0.073*	0.311 ± 0.215	0.250 ± 0.202*	5.440 ± 2.211	4.975 ± 2.272*

*p< 0.05, fine-tuned SAM_{1024} vs. fine-tuned MedSAM. 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.

Human Annotation Study

To test SAM's potential in enhancing clinical workflow, we designed a human annotation study using SAM_{112}. In the study, we observed that AI assistance significantly improved efficiency by decreasing 50% annotation time (p <0.0001; **Figure 3a**), which remains true for both expert-level and medical student-level annotators (**Figure 3b and 3c**). The AI-assisted workflow maintained the quality of left ventricle segmentation, as measured by DSC (**Figure 3d, e, and f**). Qualitative assessments indicated an improvement in performance for inexperienced annotators, which may not be completely captured by the DSC scores (**Supplementary Figure 1**).

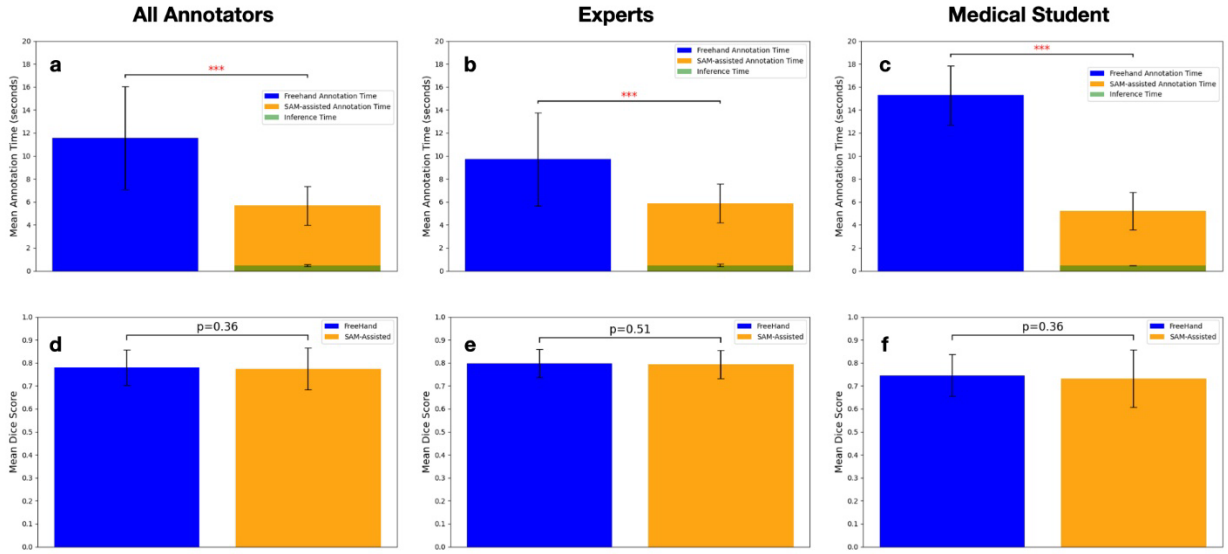

Figure 3. SAM-assisted workflow decreased the time for annotation while maintaining the quality of echocardiography image segmentation.

Discussion

The major contributions of this work include: 1) presenting a data-efficient and cost-effective strategy for training a LV segmentation model for echocardiography images based on SAM and leveraging its generalization capability for POCUS images, and 2) showing the potential of optimizing clinical

workflows through a human-in-the-loop approach, where SAM significantly improved efficiency while maintaining annotation quality.

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[3,27–29]. 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[3,25]. While the image/frame-based approach of SAM does not consider consecutive inter-frame changes and spatio-temporal features of the echocardiogram, fine-tuned SAM (both SAM_{112} and SAM_{1024}) achieved superior frame-level segmentation performance on all datasets compared to the video-based EchoNet model[3]. Importantly, there were cases with suboptimal image quality and imperfect human labels (**Figure 1**) in the EchoNet-dynamic data set[3,30], which can limit the model performance.

In terms of generalization capability, the fine-tuned SAM_{112} model demonstrated robust performance on unseen TTE (including the A2C view in CAMUS) and POCUS data, with about a 1% and 5% drop in performance, respectively. In contrast, there was a significant drop in EchoNet and UNet's performance on the POCUS dataset (**Table 1**). Similarly, when evaluated by border-sensitive metrics such as HD and ASD, foundation models (SAM and MedSAM) were able to generate more consistent borders compared to EchoNet and U-Net (**Supplementary Table 3**). This again demonstrated the advantage of leveraging the generalization capabilities of foundation models [23] in building AI solutions for the rapidly growing use of POCUS in cardiac imaging.[34] While evaluation on a larger POCUS dataset 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[34–37].

On a head-to-head comparison, fine-tuned SAM_{112} outperformed base MedSAM across almost all datasets when evaluated by DSC and IoU (**Table 1)**. This is likely due to the fact that cardiac ultrasound images were less represented in MedSAM's training set of over 1 million images. When fine-tuned under the same setting, SAM_{1024} and fine-tuned MedSAM can achieve the same level of performance and generalization capability on a specialized LV segmentation task (**Table 4**). Our study supports a data-efficient foundation model training strategy that uses representative examples for specific tasks in medical subspecialties, providing a generalizable solution that overcomes the challenge of large-scale, high-quality dataset collection in this field[19,31]. Importantly, starting with SAM also saved computational resources for training on 20 A100 (80G) GPUs in MedSAM's training process[26].

SAM's interactive capability allows the potential integration of SAM into research or clinical workflows in echocardiography labs to create high-quality segmentation masks at scales [23,25]. We demonstrated the potential of a SAM-based, human-in-loop workflow for LV segmentation, significantly reducing the annotation time and maintaining the segmentation quality across users with different experience levels (**Figure 3; Supplementary Figure 1**). Compared to fully automatic models[3], an interactive approach ensures that segmentations can be directly prompted or modified to the level of interpreters' satisfaction[25,32]. This process also has the potential to gain the trust of human users to facilitate integration[33]. 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 predict segmentations accurately [25].

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 have not specifically explored this approach in the current paper. Since this manuscript focuses on exploring effective strategies for adapting a generic foundation model like SAM, we did not modify the architecture, and as such, the study does not include substantial technical innovations. The balance of performance and generalization capability was not comprehensively explored. Additionally, how SAM can be integrated into the current echocardiography lab workflow and its real-world effects will need to be validated in a prospective setting, which is beyond the scope of the study.

In conclusion, fine-tuning SAM on a relatively small yet representative echocardiography dataset resulted in superior performance compared to the state-of-the-art echocardiography segmentation model, EchoNet, as well as base MedSAM, and achieved the same performance as fine-tuned MedSAM. We demonstrated a data-efficient and cost-effective approach for handling specialized complex echocardiography images, and SAM's potential to enhance current echocardiography workflows by improving the efficiency with a human-in-the-loop design.

Methods

Population and Data Curation

EchoNet-Dynamic: 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]. 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]. Patient characteristics were not available for this public dataset. In this study, the cases without ground truth labels were excluded from this analysis (5 from the train set, 1 from the test set) (**Supplementary Table 3**).

Mayo Clinic: A dataset from the 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 an external validation dataset. The A4C videos were reviewed by a clinical sonographer (RW) and a cardiologist (JMF). Fifty-two (33 TTE and 19 POCUS) out of the total 99 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). 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 A4C 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.

CAMUS: 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 [40]. The CAMUS cohort (n=500) had a mean age of 65.1 ± 14.4 years, 66% male, and a mean LVEF of 44.4 ± 11.9%.

Segment Anything Model (SAM), MedSAM, and U-Net

The Segment Anything Model (SAM) is an image segmentation foundation model trained on a dataset of 11 million images and 1.1 billion masks[32]. 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 [32]. 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 [32]. MedSAM is a fine-tuned version of the base SAM, trained on over 1.5 million medical image-mask pairs across 10 modalities and 30 cancer types, with trained weights available [26]. *U-Net* architecture was selected as the representative of commonly used segmentation deep learning models. In our study, the U-Net architecture[41] leveraged a ResNet-18 backbone[42].

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 by 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."[32]

Zero-shot Evaluation

The original SAM ViT-base model (model type "vit_b", checkpoint "sam_vit_b_01ec64.pth ") without modification or fine-tuning was used to segment the dataset, referred to as zero-shot learning in machine learning literature[26,32]. 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[32]. Bounding box coordinates of each left ventricle segmentation tracing were generated from the ground truth segmentations and used as the prompt for SAM[24]. Of note, while SAM supports both the bounding box and point prompts, our study focused exclusively on the bounding box prompt due to the incomplete and relatively weak borders in echocardiography images. Additionally, since the left atrium and LV are connected structures, point prompts may result in masks that incorrectly link the two chambers.

Model Fine-tuning

The SAM ViT-base model was used for fine-tuning with a procedure described by Ma et al [38]. We used the training set cases (n=7,460) of the EchoNet-Dynamic as our customized dataset without 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[38,43]. Adam optimizer[44] was used (weight decay = 0), with an initial learning rate of 2e-5 (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. The fine-tuning procedure of U-Net followed the same procedure except for using bounding boxes, and the training epoch was 23 for the best-performing model. No data augmentation was performed.

To ensure fair comparison between SAM and MedSAM, we used MedSAM's fine-tuning setup with 1024 × 1024 images, and both models were trained for 70 epochs. The two versions of fine-tuned SAM models were referred to as SAM_{112} and SAM_{1024}, respectively. The training time for SAM_{112} and SAM_{1024} was about 25 hours and 40 hours, respectively. MedSAM required a training time of about 50 hours. UNet required 16 mins.

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) [40] and a Mayo Clinic dataset including the A4C view of cases of TTE and POCUS devices.

Human Annotation Study

We conducted a human annotation study to evaluate the enhancement of segmentation efficiency and quality when utilizing a fine-tuned SAM. Three annotators (two experts and one medical student) independently annotated 100 echocardiography images, randomly selected from the EchoNet-Dynamic test set. Each one annotated 100 images with and without SAM assistance. In the SAM-assisted workflow, annotators were instructed to draw minimal bounding boxes to prompt the model. In the regular workflow, annotators manually trace the endocardium contour, as per standard clinical practice. Annotators were allowed to refine their annotations until they were satisfactory. We recorded and compared the time required to complete annotations and the quality of annotations (DSC against ground truth) using both workflows.

Statistical Model Performance Evaluation

The model segmentation performance was directly evaluated by Intersection over Union (IoU), Dice similarity coefficient (DSC), Hausdorff Distance (HD), and Average Surface Distance (ASD) against human ground truth labels[45]. IoU and DSC assess the overlap of segmented areas, while HD and ASD evaluate the alignment of segmentation borders. The formulas are provided in **Supplementary Note 1**. Depending on the normality of distribution, two-tailed, paired t-tests or Wilcoxon tests were conducted to assess the statistical significance of the differences in accuracy between models (zero-shot vs. fine-tuned SAM, EchoNet vs. fine-tuned SAM, and MedSAM vs. fine-tuned SAM), as well as the annotation time and quality metrics in the human annotation study (AI-assisted vs. non-AI-assisted), with $p<0.05$ as significant.

Ethical review and approval

All the studies have been performed in accordance with the Declaration of Helsinki. EchoNet-Dynamic dataset contains a publicly available, de-identified dataset that was approved by Stanford University Institutional Review Board and data privacy review through a standardized workflow by the Center for Artificial Intelligence in Medicine and Imaging (AIMI) and the University Privacy Office. The CAMUS dataset is also publicly available, under the approval of the University Hospital of St Etienne (France) ethical committee after full anonymization. The use of the Mayo Clinic dataset was approved by the institutional review board (protocol#22-010944); only patients providing informed consent for minimal-risk retrospective studies were included, thus the requirement for additional informed consent was waived.

Data Availability:
The EchoNet and CAMUS datasets are publicly available. The Mayo Clinic dataset can not be made publicly available due to patient privacy regulations; it is available from the corresponding author upon reasonable request.

Code Availability:
We released a checkpoint of the fine-tuned SAM and MedSAM models, along with the fine-tuning, inference, and statistical analysis code. The code and checkpoint are available on GitHub: https://github.com/chiehjuchao/SAM-Echo.git

Acknowledgment:
We would like to acknowledge Owen Crystal, MS, for his contributions to the organization of EchoNet-related data, which significantly facilitated the analysis. The authors declare no funding support for this work.

Author Contributions:
C.J.C.: Conceptualization, methodology, software, validation, data curation, formal analysis, writing – original draft, visualization.
Y.R.G.: Methodology, software.
W.K.: Software, data curation.
T.X.: Methodology, software.
L.A.: Methodology, software.
J.W.: Methodology, software.
J.M.F.: Data curation.
R.W.: Data curation.
J.J.: Data curation, software.
R.A.: Data curation, resources, writing – review and editing.
G.C.K.: Data curation, resources, writing – review and editing.
J.K.O.: Data curation, resources, writing – review and editing.
C.P.L.: Resources, writing – review and editing.
I.B.: Conceptualization, resources, writing – review and editing, supervision.
F.F.L.: Conceptualization, methodology, resources, writing – review and editing, supervision.

E.A.: Conceptualization, methodology, resources, writing – review and editing, supervision.
All authors have read and approved the manuscript.

Competing Interests:
The authors declare no competing interests.

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

References

1. Antoine, C. *et al.* Clinical Outcome of Degenerative Mitral Regurgitation. *Circulation* 138, 1317–1326 (2018).
2. Matulevicius, S. A. et al. Appropriate Use and Clinical Impact of Transthoracic Echocardiography. *JAMA Intern Med* 173, 1600–1607 (2013).
3. Ouyang, D. et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 580, 252–256 (2020).
4. Lang, R. M. 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* 28, 1-39.e14 (2015).
5. Liu, J. et al. Contemporary Role of Echocardiography for Clinical Decision Making in Patients During and After Cancer Therapy. *JACC Cardiovasc Imaging* 11, 1122–1131 (2018).
6. Fonseca, R., Jose, K. & Marwick, T. H. Understanding decision-making in cardiac imaging: determinants of appropriate use. *European Hear J - Cardiovasc Imaging* 19, 262–268 (2017).
7. Otto, C. M. 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) doi:10.1016/j.jacc.2020.11.018.
8. Tam, J. W., Nichol, J., MacDiarmid, A. L., Lazarow, N. & Wolfe, K. What Is the Real Clinical Utility of Echocardiography? A Prospective Observational Study. J Am Soc Echocardiogr 12, 689–697 (1999).
9. Pellikka, P. A. 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* 1, e181456–e181456 (2018).
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 44, 1030–1035 (2004).
11. 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).
12. Koh, A. S. et al. A comprehensive population-based characterization of heart failure with mid-range ejection fraction. *Eur J Heart Fail* 19, 1624–1634 (2017).
13. Gilbert, A. et al. Generating Synthetic Labeled Data From Existing Anatomical Models: An Example With Echocardiography Segmentation. *IEEET Med Imaging* 40, 2783–2794 (2021).
14. Salte, I. M. et al. Artificial Intelligence for Automatic Measurement of Left Ventricular Strain in Echocardiography. Jacc Cardiovasc Imaging 14, 1918–1928 (2021).
15. Huang, H. et al. Segmentation of Echocardiography Based on Deep Learning Model. *Electronics* 11, 1714 (2022).
16. Amer, A., Ye, X. & Janan, F. ResDUnet: A Deep Learning-Based Left Ventricle Segmentation Method for Echocardiography. *IEEE Access* 9, 159755–159763 (2021).
17. Jafari, M. H., Woudenberg, N. V., Luong, C., Abolmaesumi, P. & Tsang, T. Deep Bayesian Image Segmentation For A More Robust Ejection Fraction Estimation. *2021 IEEE 18th Int. Symp. Biomed. Imaging (ISBI)* 00, 1264–1268 (2021).
18. Mokhtari, M., Tsang, T., Abolmaesumi, P. & Liao, R. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. *Lect. Notes Comput. Sci.* 360–369 (2022) doi:10.1007/978-3-031-16440-8_35.
19. Yu, Y. et al. Techniques and Challenges of Image Segmentation: A Review. *Electronics* 12, 1199 (2023).
20. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* (2020) doi:10.48550/arxiv.2010.11929.
21. Vaswani, A. et al. Attention Is All You Need. *arXiv* (2017) doi:10.48550/arxiv.1706.03762.
22. Liu, Y. et al. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. *arXiv* (2023).
23. Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. *arXiv* (2021) doi:10.48550/arxiv.2108.07258.
24. Kirillov, A. et al. Segment Anything. *arXiv* (2023).
25. Mazurowski, M. A. et al. Segment Anything Model for Medical Image Analysis: an Experimental Study. *arXiv* (2023).
26. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* 15, 654 (2024).
27. Mintz, G. S. & Kotler, M. N. Clinical Value and Limitations of Echocardiography: Its Use in the Study of Patients With Infectious Endocarditis. *Arch. Intern. Med.* 140, 1022–1027 (1980).
28. Mondillo, S., Maccherini, M. & Galderisi, M. Usefulness and limitations of transthoracic

echocardiography in heart transplantation recipients. *Cardiovasc. Ultrasound* 6, 2–2 (2008).
29. Abdulla, A. M., Frank, M. J., Canedo, M. I. & Stefadouros, M. A. Limitations of echocardiography in the assessment of left ventricular size and function in aortic regurgitation. *Circulation* 61, 148–155 (2018).
30. Tromp, J. et al. Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. *Lancet Digital Heal* 4, e46–e54 (2022).
31. Zhang, D. et al. Deep Learning for Medical Image Segmentation: Tricks, Challenges and Future Directions. *arXiv* (2022) doi:10.48550/arxiv.2209.10307.
32. Kirillov, A. et al. Segment Anything. *arXiv* (2023) doi:10.48550/arxiv.2304.02643.
33. Bao, Y., Cheng, X., Vreede, T. D. & Vreede, G.-J. D. Investigating the relationship between AI and trust in human-AI collaboration. *Proc. 54th Hawaii Int. Conf. Syst. Sci.* (2021) doi:10.24251/hicss.2021.074.
34. Huang, G. S., Alviar, C. L., Wiley, B. M. & Kwon, Y. The Era of Point-of-Care Ultrasound Has Arrived: Are Cardiologists Ready? *Am J Cardiol* 132, 173–175 (2020).
35. Kalagara, H. et al. Point-of-Care Ultrasound (POCUS) for the Cardiothoracic Anesthesiologist. *J Cardiothor Vasc An* 36, 1132–1147 (2022).
36. Palmero, S. L. et al. Point-of-Care Ultrasound (POCUS) as an Extension of the Physical Examination in Patients with Bacteremia or Candidemia. *J Clin Medicine* 11, 3636 (2022).
37. 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* 19, 39 (2021).
38. Ma, J. & Wang, B. Segment Anything in Medical Images. *arXiv* (2023) doi:10.48550/arxiv.2304.12306.
39. Wu, J. et al. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *arXiv* (2023).
40. Leclerc, S. et al. Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. *IEEE Trans. Méd. Imaging* 38, 2198–2210 (2019).
41. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* (2015) doi:10.48550/arxiv.1505.04597.
42. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conf Comput Vis Pattern Recognit Cvpr* 770–778 (2016) doi:10.1109/cvpr.2016.90.
43. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18, 203–211 (2021).
44. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* (2014) doi:10.48550/arxiv.1412.6980.
45. Maier-Hein, L. et al. Metrics reloaded: Recommendations for image analysis validation. *arXiv* (2022) doi:10.48550/arxiv.2206.01653.

Figure Legends

Figure 1. 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 endocardial 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 2. 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 the anticipated left ventricular geometry on visualization. Note that POCUS images generally had worse quality compared to transthoracic echocardiography images.

Figure 3. SAM-assisted workflow decreased the time for annotation while maintaining the quality of echocardiography image segmentation. Panel a. With SAM's assistance, the average annotation time significantly decreased by 50%. Panel b. Among experienced annotators, the workflow decreased annotation time by 39.2%; Panel c. It also reduced 66.0% annotation time for an inexperienced annotator. Panel d, e, and f. No significant difference was observed in segmentation quality when measured by the Dice Similarity Score. ***$p<0.0001$.