MemSAM: Taming Segment Anything Model for Echocardiography Video Segmentation

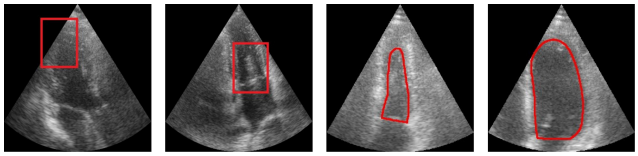
Xiaolong Deng[1], Huisi Wu[1]*, Runhao Zeng[2], Jing Qin[3]
[1] College of Computer Science and Software Engineering, Shenzhen University
[2] College of Mechatronics and Control Engineering, Shenzhen University
[3] Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University
dengxiaolong2022@email.szu.edu.cn, {hswu, zengrh}@szu.edu.cn, harry.qin@polyu.edu.hk

Abstract

We propose a novel echocardiographical video segmentation model by adapting SAM to medical videos to address some long-standing challenges in ultrasound video segmentation, including (1) massive speckle noise and artifacts, (2) extremely ambiguous boundaries, and (3) large variations of targeting objects across frames. The core technique of our model is a temporal-aware and noise-resilient prompting scheme. Specifically, we employ a space-time memory that contains both spatial and temporal information to prompt the segmentation of current frame, and thus we call the proposed model as MemSAM. *In prompting, the memory carrying temporal cues sequentially prompt the video segmentation frame by frame. Meanwhile, as the memory prompt propagates high-level features, it avoids the issue of misidentification caused by mask propagation and improves representation consistency. To address the challenge of speckle noise, we further propose a memory reinforcement mechanism, which leverages predicted masks to improve the quality of the memory before storing it. We extensively evaluate our method on two public datasets and demonstrate state-of-the-art performance compared to existing models. Particularly, our model achieves comparable performance with fully supervised approaches with limited annotations. Codes are available at https://github.com/dengxl0520/MemSAM.*

1. Introduction

Cardiovascular diseases are the leading cause of mortality worldwide according to statistics of the World Health Organization (WHO) [6]. Echocardiography is an important yet unique tool for assessing cardiovascular function. Due to its portability, low cost, and real-time nature, echocardiography is commonly used as the first-line examination method in clinical practice [1]. However, echocardiography usually requires manual evaluation by experienced physicians, and the quality of assessment heavily relies on physicians' expertise [17, 19]. To the end, there are often large inter- and intra-observer differences among manual assessments [16]. In addition, the assessment requires manual tracking of ventricular size, which is laborious, time-consuming, and error-prone. In this regard, automated assessment approaches are highly demanded in clinical practice.



(a) (b) (c) (d)

Figure 1. The challenges of echocardiography video segmentation: (a) blurred contours, (b) speckle noise, and (c-d) the change of scale across frames (two frames of the same video).

Echocardiographic assessment and diagnosis are usually based on the interpretation of ejection fraction and ventricular volumes [4], which requires accurate segmentation of key structures from echocardiographic videos such as the left ventricular endocardium. However, automatic echocardiographic segmentation has always been a challenging task. First, as shown in Figure 1 (a,b), due to limitations of ultrasound imaging, there are a lot of adverse factors affecting the quality of echocardiographic videos, such as low signal-to-noise ratio, speckle noise, edge dropout, and shadows caused by structures like dense muscle and ribs, making it difficult to identify the boundaries of key anatomical structures [3, 30]. Second, shape and scale variations of cardiac structures are large within and between videos (see Figure 1 (c,d)). Finally, annotation of echocardiographic videos is labor-intensive and time-consuming, and hence physicians usually only annotate end-systole and end-

*Corresponding Author

diastole frames. To the end, we have to segment echocardiographic videos with limited and sparse annotations.

In recent years, many deep learning methods have been proposed for echocardiographical video segmentation [32, 33, 35, 36], but they still cannot achieve satisfactory results due to the low quality of ultrasound videos and limited annotations. Recently, a large vision model, Segment Anything Model (SAM) [15] has been proposed and achieved significant success in many natural image segmentation tasks. Some researchers have endeavored to adapt it to medical image segmentation tasks in order to take advantage of SAM's powerful representation capability to alleviate inadequate training samples. However, most of these studies focus on 2D images and how to adapt SAM for medical video segmentation remains an unexplored and challenging task. Applying SAM directly to videos would ignore temporal clues, and may result in temporally inconsistent segmentation [26, 34]. For example, as shown in Figure 1, fast-changing echocardiographical videos have obvious temporal discontinuities in the shape and scale of targeting objects. In addition, ambiguous boundaries caused by massive speckle noise and artifacts will greatly prohibit SAM from unleashing its representation capability.

In this paper, we propose a novel echocardiographical video segmentation model by adapting SAM to medical videos, which have some unique characteristics compared with natural videos. The core technique of our model is a temporal-aware and noise-resilient prompting scheme. Specifically, we employ a space-time memory that contains both spatial and temporal information to prompt the segmentation of current frame, and thus we call the proposed model as *MemSAM*. In prompting, the memory carrying temporal cues sequentially prompt the video segmentation frame by frame. Meanwhile, as the memory prompt propagates high-level features, it avoids the issue of misidentification caused by mask propagation and improves representation consistency. To address the challenge of speckle noise, we propose a memory reinforcement mechanism, which leverages predicted masks to improve the memory quality before storing it. We build our model on SAMUS [18], a medical foundation model based on SAM, which enables our model to be more adaptable to medical data. Finally, we conducted extensive experiments on two publicly available datasets. Our contributions can be summarized as follows:

- We propose a novel echocardiography video segmentation model based on SAM. The core component of our model is a new prompting approach, which is able to provide both spatial and temporal cues to improve representation consistency and segmentation accuracy.
- We further propose the memory reinforcement module to enhance the memory before storing it, thereby alleviating the adverse effects of speckle noise and motion artifacts during the memory prompting.
- We extensively evaluate our method on two public datasets and demonstrate state-of-the-art performance compared to existing models. Particularly, our model achieves comparable performance with fully supervised approaches with limited annotations.

2. Related Work

2.1. SAM in Medical Image Segmentation

The SAM demonstrates excellent zero-shot generalization capabilities when applied to natural images [11, 23]. However, due to the complex shapes, blurred boundaries, and significant scale variations inherent in medical images, it still falls short of being directly applicable to medical image segmentation [14]. Some general works have attempted to adapt SAM from natural images to medical images [42], such as MedSAM [21], MSA [37] and SAMed [40]. MedSAM does not change the SAM network structure, but uses bounding box prompts more suitable for the medical domain, and focuses on fine-tuning the mask decoder. The purpose of MSA and SAMed is to modify the Image Encoder to adapt to medical images. MSA realizes this by adding an adapter to the Image Encoder. SAMed uses a strategy based on low-rank (LoRA) strategy to fine-tune the Image Encoder. In more specialized domains, SAMUS [18] and SonoSAM [28] focused on ultrasound images have also been proposed. Among them, SAMUS adapts better to ultrasound images by adding adapters and additional CNN branches. SonoSAM uses knowledge distillation to extract specific knowledge from medical images. However, these methods are limited to image segmentation and have yet to be extended to video data, relying heavily on dense annotations and prompts to attain adequate performance. In contrast, the aim of this work is to investigate leveraging temporal cues within video to enable model training with only sparse annotations and minimal prompts.

2.2. SAM in Video Segmentation

Although the extension of SAM to the video domain remains relatively underexplored, some preliminary works have been proposed to address natural video segmentation tasks. A common approach involves integrating SAM with prevalent video segmentation architectures, as exemplified by SAM-Track [9] and TAM [38]. SAM-Track uses SAM to obtain keyframe segments as references, then leverages DeAOT [39] to propagate the reference frames throughout the video sequence. TAM combines SAM and XMem [7] by initially generating coarse masks with SAM and weak prompts, then employing XMem for continued tracking. When segmentation quality declines, TAM refines the SAM outputs using XMem's prediction probabilities and affinities as prompts. More recently, SAM-PT [27] introduced a unique point-tracking technique to generate masks and

track objects. However, these methods are only suitable for relatively simple natural image scenes and are difficult to apply to medical image segmentation. For example, for complex, dynamic, and noisy ultrasound images, the intermediate features of XMem will carry noise that incorrectly prompts SAM. Furthermore, when XMem passes intermediate parameters to SAM, converting them to mask prompts loses the higher-level semantics of the original features. Similar to TAM, our method is also based on XMem. Critically, rather than a naive combination, our method emphasizes maintaining semantic consistency during feature transfer and mitigating pervasive background noise in medical imaging data.

2.3. Space-Time Memory Methods

Mainstream video temporal modeling methods include multi-frame aggregation and space-time memory networks. Multi-frame aggregation learns temporal features by aggregating semantic information from adjacent frames. In comparison, space-time memory models video temporal information by propagating semantic information along the temporal dimensions. Although multi-frame aggregation is widely used, its GPU memory requirement increases rapidly with video length, limiting its application in long video processing. In contrast, space-time memory networks can significantly reduce memory consumption while ensuring temporal modeling, making them more suitable for extension to areas like medical video analysis. Space-time memory networks (STM) were first proposed by Oh et al. [24] for video object segmentation tasks. Subsequent methods including STCN [8], XMem [7], and XMem++ [2] have demonstrated immense potential for general video segmentation. However, these methods require an annotated reference keyframe for the videos to be segmented, which is difficult for our task.

3. Method

SAM is a powerful prompt-based segmentation framework that utilizes prompts to track targets to segment after learning good representations [31, 41]. How to properly prompt is an issue worth studying. Existing SAM and its variants perform well on image segmentation, including natural and medical images. However, they cannot utilize temporal clues in videos when directly migrated to video segmentation, ignoring the spatio-temporal consistency in videos. Moreover, applying them directly to videos would require prompting every frame, which is inelegant and redundant for video segmentation. We aim to design a memory prompting method to extend the SAM framework and avoid prompting every frame in a video. Meanwhile, annotating every frame in a video is also extremely difficult, especially for echocardiograms where it is hard to acquire abundant annotations. Therefore, a method that can accomplish semi-supervised tasks is needed.

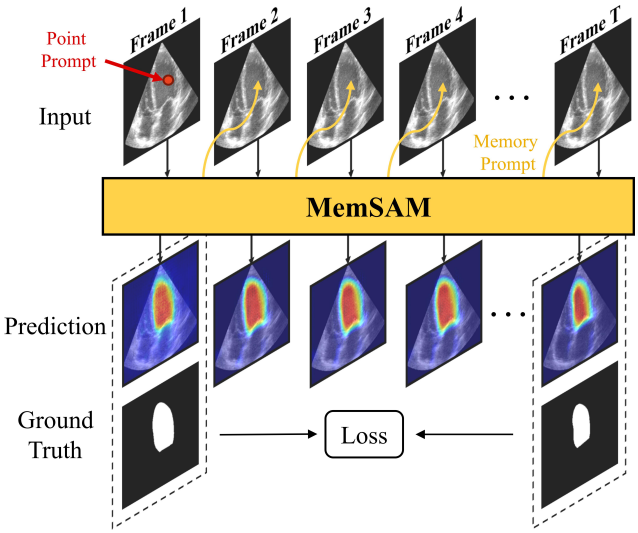


Figure 2. The workflow of MemSAM, in which only the first frame of the video uses the simplest positive point prompt (red arrow), and subsequent frames use memory prompts (yellow arrow). Finally, the loss is calculated for the Prediction and Ground Truth of the supervised frame.

Therefore, we propose MemSAM for solving semi-supervised problems with fewer annotations and prompts in echocardiography. The proposed MemSAM framework processes videos in a sequential frame-by-frame manner, as illustrated in Figure 2. Each input video of length T frames is fed into the MemSAM model one frame at a time. Initially, randomly sampled points in the foreground are provided as prompts for the first frame to guide the model. For subsequent frames, MemSAM relies solely on memory prompts rather than external prompts. After prediction by MemSAM, the prediction of supervised frames will calculate loss with the Ground Truth.

3.1. Overview

Figure 3 shows more details inside the MemSAM framework. The MemSAM mainly consists of two components, the SAM component and the Memory component. The SAM component adopts an architecture identical to the original SAM, composed of an image encoder, a prompt encoder, and a mask decoder. The image encoder employs the Vision Transformer (ViT) [12] as the backbone to encode input images into image embedding E_i. The prompt encoder ingests external prompts, such as point prompts, and encodes them into a c-dimensional embedding. Subsequently, the mask decoder integrates the image and prompt embeddings to predict segmentation masks.

Among them, image embedding is mapped to the memory feature space through the projection layer, and then

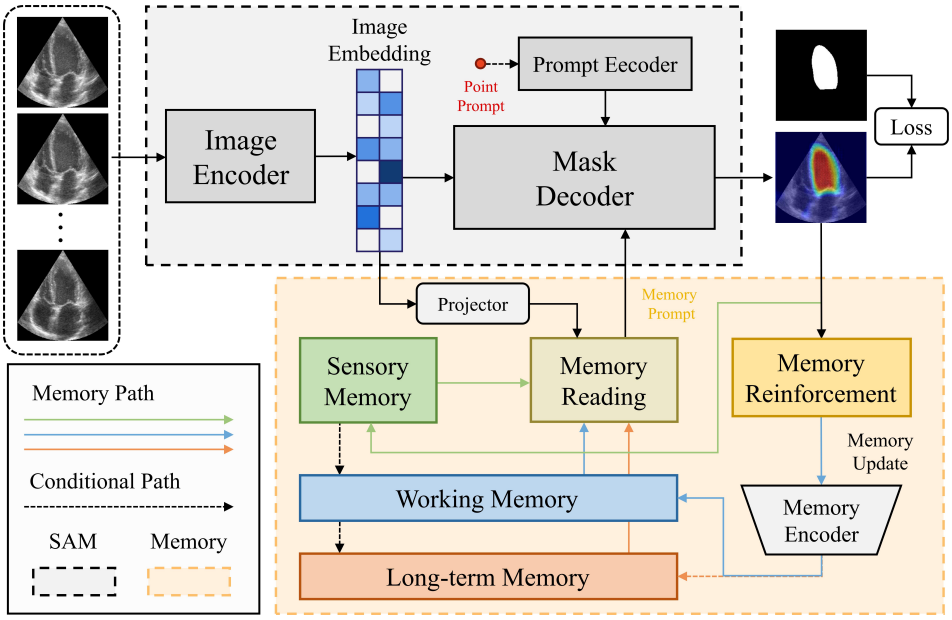


Figure 3. Overview of our MemSAM, which consists of SAM and Memory. The input image is first encoded into an image embedding by the image encoder of SAM. After obtaining the point prompt and memory prompt, the mask is output from the mask decoder.

we perform memory reading to obtain the memory prompt from multiple feature memory (sensory memory, working memory, and long-term memory) and provide it to the mask decoder. Finally, after passing the memory reinforcement and memory encoder, the memory will be updated. Detailed information on important components is provided below.

3.2. Memory Reading

The Memory Reading block in Figure 4 shows the process of generating memory embedding E_m from image embedding E_i, which is input to the mask decoder as a memory prompt. The image embedding E_i^t of frame t is projected through a projection layer to generate the query q^t. This query q^t is then used to perform an affinity query against the memory keys and values to obtain the readout features F^t. The process can be formulated as:

$$F^t = v^{t-1} \cdot W(k^{t-1}, q^t) \tag{1}$$

where $k^{t-1} = k_w^{t-1} \oplus k_{lt}^{t-1}$ and $v^{t-1} = v_w^{t-1} \oplus v_{lt}^{t-1}$. The $\oplus$ denotes concatenation and superscripts 'w' and 'lt' denote working and long-term memory respectively. The $W(k^{t-1}, q^t)$ represents the affinity matrix between the query q^t and the memory key k^{t-1}, and it captures the correlation between q^t and k^{t-1}. It can be obtained by computing the similarities between q^t and k^{t-1}, followed by normalization. The specific computation process can be formulated as follows:

$$W(k^{t-1}, q^t) = softmax(S(k^{t-1}, q^t)) \tag{2}$$

where S is the similarity calculation. In order to encode the confidence level of memory elements and focus on more important channels, we adopt *anisotropic L2 similarity* [7] as the similarity function. Finally, the readout feature F^t is fused with the sensory memory h^{t-1} and E_i^t to obtain the memory embedding E_m^t, which is formulated as:

$$E_m^t = Fusion(E_i^t, F^t \oplus h^{t-1}) \tag{3}$$

3.3. Memory Reinforcement

Compared with natural images, ultrasound images contain more complex noise, which means the image embedding generated by the image encoder will inevitably carry noise. If the noisy features are updated to the memory without any processing, it may lead to the accumulation and propagation of errors. To mitigate the influence of noise on memory updates, we employ a Memory Reinforcement module to enhance the discriminability of the feature representations in memory. As shown in the memory reinforcement in Figure 4, we reinforce the memory with the segmentation results before the memory update, aiming to emphasize foreground features and reduce the impact of background noise.

Specifically, for the probability map $P^t \in \mathbb{R}^{B\times1\times H\times W}$ output by the mask decoder, we first *downsample* it into a $P_d^t \in \mathbb{R}^{B\times1\times h\times w}$ of the same size as the image feature $E_i^t \in \mathbb{R}^{B\times C\times h\times w}$, and then concat it with E_i^t along the channel dimension to obtain $F^t \in \mathbb{R}^{B\times(C+1)\times h\times w}$. We use a convolutional layer $Conv_{3\times3}$ with a convolution kernel size of 3×3 to process F, in order to limit the receptive

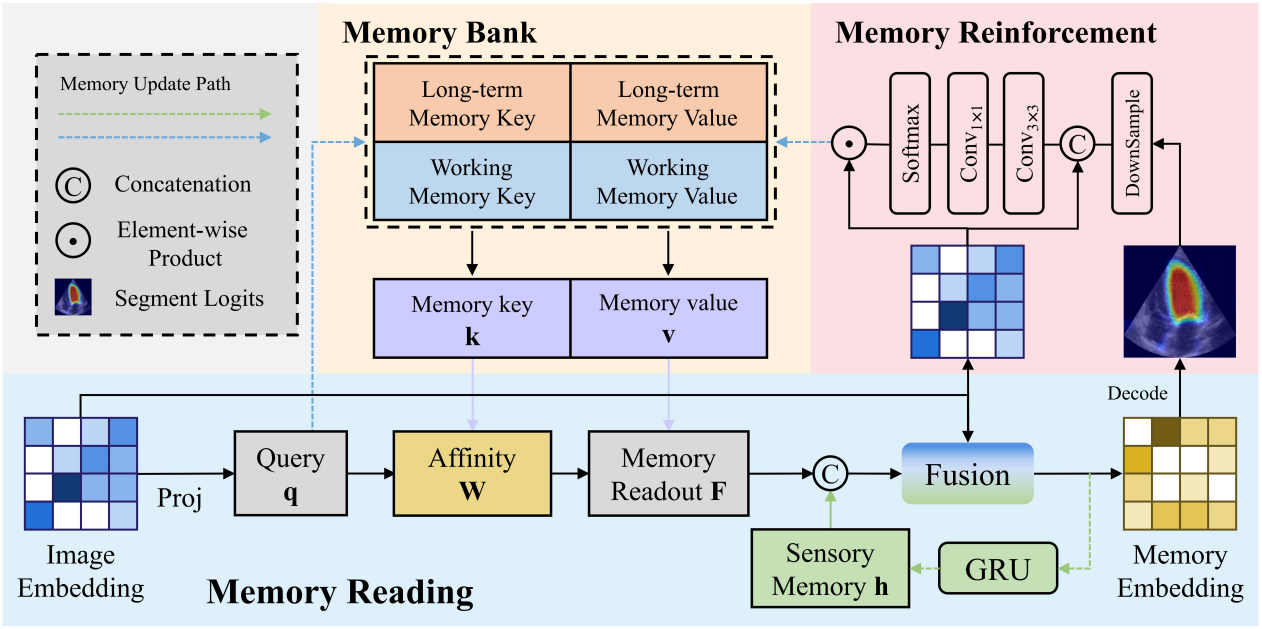


Figure 4. More details of Memory Reading and Memory Reinforcement. The Memory Update process is represented briefly.

field of each pixel. This process generates a local attention weight feature $F_w^t \in \mathbb{R}^{B\times C_{mid}\times h\times w}$. Then we use a $Conv_{1\times1}$ convolution layer to change the number of channels of F_w^t to obtain $F_w^t \in \mathbb{R}^{B\times C\times h\times w}$, and finally calculate the output features:

$$F_o^t = P_d^t \odot softmax(F_w^t) \tag{4}$$

where $\odot$ represents the element-wise product. F_o will finally be inserted into the Working Memory value. Through this mechanism, we use segmentation results to maintain foreground features, weaken the impact of background noise on memory updates, and enhance the distinguishability of feature expressions in memory.

3.4. Memory Update

The memory to be updated includes memory bank and sensory memory. memory bank further consists of working memory and long-term memory, where long-term memory is only utilized for long videos and is omitted here. The sensory memory h^t is updated by:

$$h^t = GRU(h^{t-1} \oplus E_m^t) \tag{5}$$

where GRU is Gated Recurrent Unit [10]. The key k_w^t and value v_w^t of working memory are updated as follows:

$$k_w^t = q^t,\ v_w^t = v_w^{t-1} \oplus F_o^t \tag{6}$$

4. Experiment

4.1. Datasets and Evaluation Metrics

We evaluated our approach on two widely used publicly available echocardiography datasets, CAMUS [16] and EchoNet-Dynamic [25].

CAMUS Dataset contains 500 cases, which include 2D apical two-chamber and apical four-chamber view video. CAMUS provides annotations across all frames.

EchoNet-Dynamic Dataset contains 10,030 2D apical two-chamber view videos. Each video provides the area of the left ventricle in the form of an integral. Only label end-systolic and end-diastolic phases.

To comprehensively assess the effectiveness of our method in semi-supervised video segmentation, the CAMUS dataset was adapted into two variants: CAMUS-Full and CAMUS-Semi. CAMUS-Full utilizes annotations for all frames during training, whereas CAMUS-Semi only uses annotations for the end-diastolic (ED) and end-systolic (ES) frames. During testing, both datasets were evaluated using complete annotations. We uniformly sampled videos from the dataset, cropping them to 10 frames each. The cropping ensured that the ED frame is the first frame, the ES frame is the last frame, and the resolution is resized to 256×256. For the CAMUS dataset, we divided it into training, validation, and test sets in a ratio of 7:1:2, while we used the original split for the EchoNet-Dynamic dataset.

We employed widely used metrics such as mean Dice coefficient (**mDice**) and mean Intersection over Union (**mIoU**) for segmentation evaluation, along with Hausdorff Distance-95% (**HD95**) and Average Symmetric Surface

Method	CAMUS-Semi				EchoNet-Dynamic			
	mDice ↑	mIoU ↑	HD95 ↓	ASSD ↓	mDice ↑	mIoU ↑	HD95 ↓	ASSD ↓
UNet [29]	90.13	82.36	5.77	2.35	91.36	83.27	4.98	3.01
SwinUNet [5]	88.84	80.33	6.10	2.60	87.79	80.14	6.61	5.71
H2Former [13]	91.31	84.30	5.27	2.05	90.21	82.46	5.12	3.78
MedSAM [21]	85.42	75.14	8.42	3.34	86.47	79.19	7.97	4.88
MSA [37]	88.03	78.98	7.53	2.85	87.91	78.34	6.67	4.34
SAMed [40]	87.45	78.14	9.17	3.10	86.35	78.96	7.12	4.59
SonoSAM [28]	89.80	81.79	6.60	2.45	89.61	82.33	6.58	3.80
SAMUS [18]	91.11	83.94	5.08	2.07	91.79	84.32	5.35	3.22
MemSAM	**93.31**±3.04	**87.61**±5.12	**3.82**±1.80	**1.57**±0.72	**92.78**±3.38	**85.89**±5.12	**4.57**±2.34	**2.71**±0.78

Table 1. Segmentation performance of the proposed method with state-of-the-art methods on the CAMUS-Semi and EchoNet-Dynamic datasets. HD95 and ASSD are measured in millimeters (mm) in CAMUS-Semi, while in pixels in EchoNet-Dynamic. Our results are expressed as mean ± standard deviation.

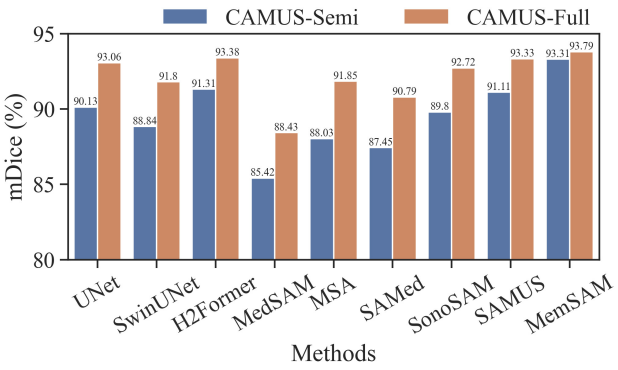


Figure 5. Segmentation performance of the proposed method with state-of-the-art methods on the CAMUS-Semi and CAMUS-Full datasets.

Method	CAMUS-Semi		
	corr (%) ↑	bias ↓	std ↓
UNet [29]	67.15	11.65	9.39
SwinUNet [5]	59.41	6.90	9.06
H2Former [13]	58.61	**0.69**	**7.49**
MedSAM [21]	41.63	11.22	11.19
MSA [37]	31.00	13.25	14.96
SAMed [40]	28.22	13.34	12.24
SonoSAM [28]	56.18	11.83	9.12
SAMUS [18]	67.55	7.02	9.16
MemSAM	**78.92**	4.86	11.10

Table 2. Clinical metrics comparison against different state-of-the-art methods on the CAMUS-Semi datasets.

Distance (**ASSD**). The standard deviations of these metrics were also reported. In addition, we also report three statistical metrics of Left Ventricular Ejection Fraction (LV_{EF}). We estimate the prediction LV_{EF} according to Simpson's biplane method of disks (SMOD), which is provided in the CAMUS dataset. Note that different implementation methods will have a significant impact on the final LV_{EF} results. SMOD estimates LV_{EF} from end diastole and end systole time instances from apical two and four chambers views. Compared with Simpson's single plane rule, SMOD's estimation solution is more accurate and reliable. For the prediction and ground truth LV_{EF}, we calculate the pearson correlation coefficient (**corr**), mean bias (**bias**), and standard error (**std**).

4.2. Implementation Details

For the SAM component, we utilize SAMUS [18], which is suitable for ultrasound images and has a more friendly deployment cost. Only the layers of the image encoder were trained, while the remaining components inherited parameters from the original SAM and were frozen. We trained for 100 epochs on the CAMUS dataset and 50 epochs on EchoNet-Dynamic. The base learning rate was set to $1e-4$, and optimization was performed using the AdamW optimizer [20]. The same loss functions (dice loss [22] and binary cross-entropy loss) as SAMUS were utilized. During the training phase, we applied gamma enhancement, random scale, random rotation, and random contrast with a probability of 0.5 each.

4.3. Comparison with State-of-the-art Methods

We extensively selected different types of comparison methods, including traditional image segmentation models and medical foundation models. The three traditional image segmentation models are respectively the CNN-based UNet [29], Transformer-based SwinUNet [5], and CNN-Transformer hybrid H2Former [13]. The medical-

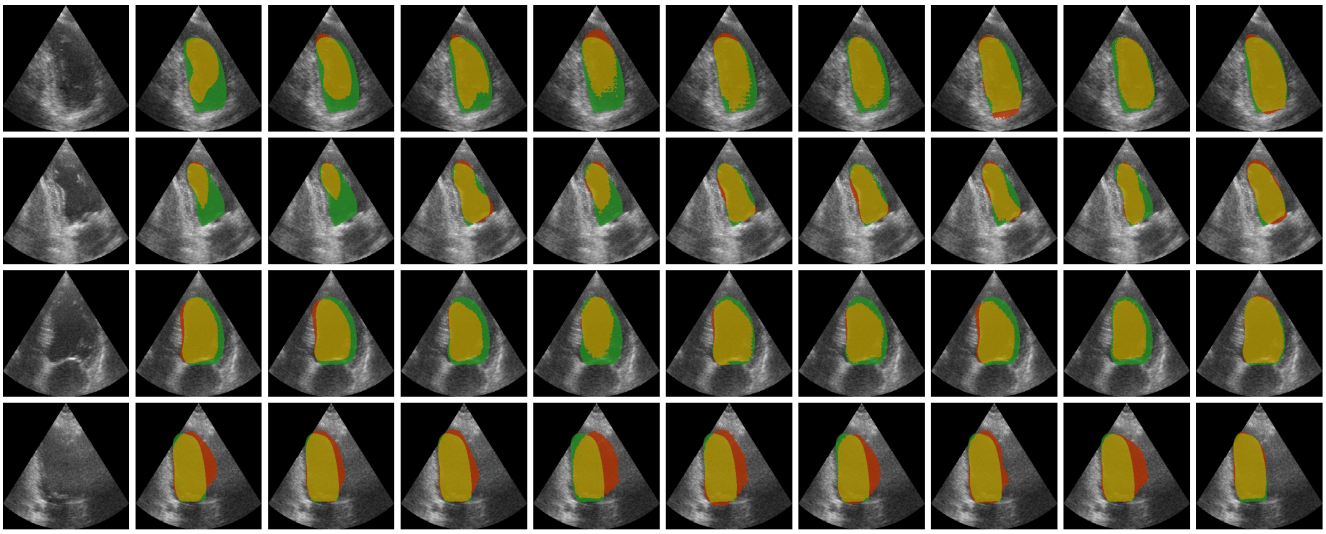


Figure 6. Visual comparison with state-of-the-art methods on the CAMUS-Semi test set. Green, red, and yellow regions represent the ground truth, prediction, and overlapping regions, respectively.

Setting	mDice	mIoU	HD95	ASSD
All components	**93.31**	**87.61**	**3.82**	**1.57**
no point prompt	93.17	87.38	4.55	1.66
no memory prompt	91.11	83.94	5.08	2.07
no memory reinforce	92.86	86.83	4.03	1.67

Table 3. Ablation study on different components of MemSAM on the CAMUS-Semi dataset.

adapted SAM models include MedSAM [21], MSA [37], SAMed [40], SonoSAM [28], and SAMUS [18]. Among them, SonoSAM and SAMUS focus on ultrasound images.

Quantitative comparison. The quantitative comparison results are shown in Table 1. Among these state-of-the-art methods, H2Former and SAMUS perform relatively well on the two datasets, benefiting from the CNN-Transformer architecture and ultrasound image optimization, respectively. However, without utilizing the temporal attributes of videos under scarce annotations, these models still lag our approach. The experiments validate that our method achieves state-of-the-art performance given limited annotations.

To further evaluate our method, we compared it under the same setting on CAMUS-Semi and CAMUS-Full datasets. The results are shown in Figure 5. It can be seen that conventional methods like UNet and H2Former, and ultrasound-specialized methods like SonoSAM and SAMUS, recover decent results given full annotations. Although our approach has marginal gains from semi-supervised to fully-supervised settings, it still outperforms other competitors under both. Notably, the medical foundation models require per-frame prompts under full supervision, while we only require a point prompt. The experiments validate that our method achieves comparable performance to full annotations with sparse labels, using far fewer external prompts.

Our method's comparison with state-of-the-art methods in LV_{EF} estimation is shown in Table 2. Under limited annotations, previous state-of-the-art methods have not been satisfactory in terms of the corr. This underperformance of state-of-the-art methods is attributable to two factors. Firstly, the segmentation accuracy of state-of-the-art methods themselves remains insufficient. Secondly, the SMOD estimation solution demands high segmentation quality, requiring both two-chamber and four-chamber views to yield accurate quantification for robust LV_{EF} evaluation.

Qualitative comparison. We present visualizations for some challenging cases. As shown in Figure 6, the images in rows 1-2 contain speckle noise around the left ventricle, which misleads some conventional and medical foundation models to incorrectly identify them as ventricle edges. Rows 3-4 contain instances with severely blurred boundaries, where almost all competitors over-segment beyond the true ventricular boundary, while our method precisely delineates the boundary. These visualizations demonstrate that our method robustly handles poor image quality cases.

4.4. Ablation Studies

To evaluate each component of MemSAM, we performed ablation studies on the CAMUS-Semi dataset.

Effectiveness of each component. We conducted ablation experiments on the main components of MemSAM to analyze the contribution of each component to our framework. We experimented with no point prompts, no memory

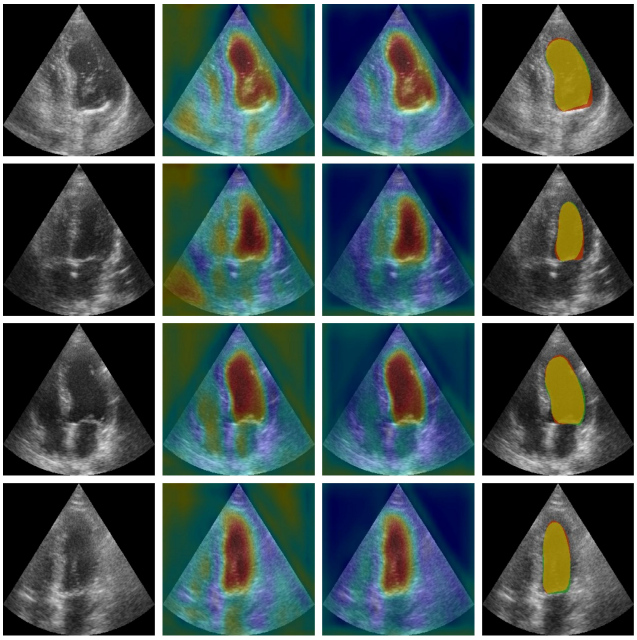


Figure 7. Feature visualization before and after memory reinforcement. After memory reinforcement, the model pays more attention to the ventricular area.

prompts, and no memory reinforcement. The experimental results are shown in the Table 3. Removing point prompts incurred only a minor performance decline. The location of the left ventricle in echocardiograms is relatively fixed, so even lacking initial frame point prompts does not greatly affect performance. Omitting memory prompts degraded performance to the SAMUS baseline. Disabling memory reinforcement resulted in a 0.45% mean Dice reduction. We visualized the memory embeddings before and after memory reinforcement. As shown in Figure 7, before reinforcement, the model attends not only to the ventricle location (deep red regions) but also to other areas (light yellow regions), which can cause incorrect activation and accumulate errors. After reinforcement, there is a greater focus on the ventricular regions, resulting in more precise segmentation.

Impact of different numbers of point prompts. To investigate the impact of different numbers of point prompts, we conducted ablation studies evaluating our proposed method with 1, 2, 3, 5, and 10 randomly sampled point prompts. Each prompt configuration was evaluated over three trials with different random seeds. The experimental results in Figure 8 demonstrate the robustness of our proposed method to numbers of randomly sampled point prompts, as evidenced by minimal variation in model performance. Since our prompt encoder is frozen, and we only use point prompts in the first frame, changes in the number of point prompts have little impact on our proposed method (only

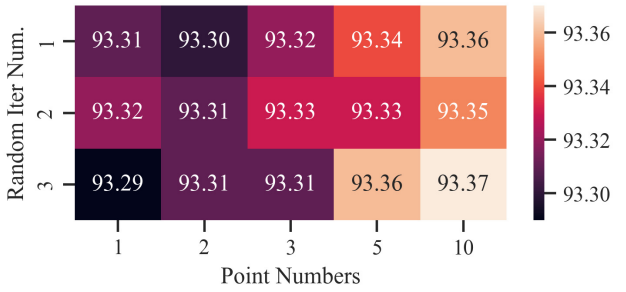


Figure 8. Ablation study of point prompts.

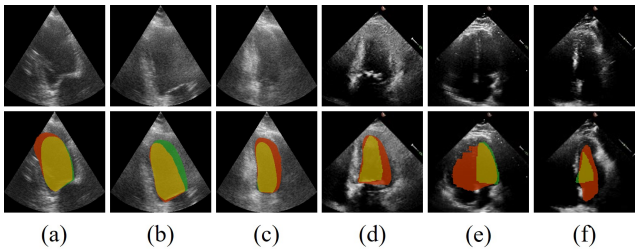


Figure 9. Failure cases on the CAMUS (a-c) and EchoNet-Dynamic (d-f) test sets.

affecting the prediction of the first frame). The experimental results show that our method requires very few point prompts to perform effectively.

5. Conclusion

In this paper, we propose a novel semi-supervised video segmentation framework for echocardiography video segmentation, aiming to effectively extend SAM to the video domain and achieve comparable performance to full supervision with limited annotations and prompting.

However, as shown in Figure 9, our method results in the entire video sequence being unable to be accurately segmented when the initial frame image is extremely poor. Future research could investigate techniques to achieve more robust initialization and test the effectiveness of our approach in more domains, and could further explore ways to reduce computational cost and lightweight the model.

Acknowledgments

This work was supported partly by National Natural Science Foundation of China (No. 62273241), Natural Science Foundation of Guangdong Province, China (Nos. 2024A1515011946 and 2023A1515011512), the Stable Support Plan Program of the Shenzhen Natural Science Foundation (No. 20220809180405001) and the Innovation and Technology Fund under Guangdong-Hong Kong Technology Cooperation Funding Scheme (ITF-TCFS) (project no. GHP/050/20SZ).

References

[1] Zeynettin Akkus, Yousof H Aly, Itzhak Z Attia, Francisco Lopez-Jimenez, Adelaide M Arruda-Olson, Patricia A Pellikka, Sorin V Pislaru, Garvan C Kane, Paul A Friedman, and Jae K Oh. Artificial intelligence (ai)-empowered echocardiography interpretation: a state-of-the-art review. *Journal of clinical medicine*, 10(7):1391, 2021. 1

[2] Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, and Hao Li. Xmem++: Production-level video segmentation from few annotated frames. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 635–644, 2023. 3

[3] Nagashettappa Biradar, Mohan Lal Dewal, and Manoj Kumar Rohit. Speckle noise reduction in b-mode echocardiographic images: A comparison. *IETE Technical Review*, 32 (6):435–453, 2015. 1

[4] Matteo Cameli, Sergio Mondillo, Marco Solari, Francesca Maria Righini, Valentina Andrei, Carla Contaldi, Eugenia De Marco, Michele Di Mauro, Roberta Esposito, Sabina Gallina, et al. Echocardiographic assessment of left ventricular systolic function: from ejection fraction to torsion. *Heart failure reviews*, 21:77–94, 2016. 1

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022. 6

[6] Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in Cardiovascular Medicine*, 7:25, 2020. 1

[7] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision*, pages 640–658. Springer, 2022. 2, 3, 4

[8] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 3

[9] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 2

[10] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 5

[11] Can Cui, Ruining Deng, Quan Liu, Tianyuan Yao, Shunxing Bao, Lucas W Remedios, Yucheng Tang, and Yuankai Huo. All-in-sam: from weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning. *arXiv preprint arXiv:2307.00290*, 2023. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[13] Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Transactions on Medical Imaging*, 2023. 6

[14] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *arXiv preprint arXiv:2304.14660*, 2023. 2

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[16] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019. 1, 5

[17] Honghe Li, Yonghuai Wang, Mingjun Qu, Peng Cao, Chaolu Feng, and Jinzhu Yang. Echoefnet: Multi-task deep learning network for automatic calculation of left ventricular ejection fraction in 2d echocardiography. *Computers in Biology and Medicine*, 156:106705, 2023. 1

[18] Xian Lin, Yangyang Xiang, Li Zhang, Xin Yang, Zengqiang Yan, and Li Yu. Samus: Adapting segment anything model for clinically-friendly and generalizable ultrasound image segmentation. *arXiv preprint arXiv:2309.06824*, 2023. 2, 6, 7

[19] Fei Liu, Kun Wang, Dan Liu, Xin Yang, and Jie Tian. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. *Medical Image Analysis*, 67:101873, 2021. 1

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[21] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2, 6, 7

[22] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 6

[23] Shentong Mo and Yapeng Tian. Av-sam: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023. 2

[24] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235, 2019. 3

[25] David Ouyang, Bryan He, Amirata Ghorbani, Matt P Lungren, Euan A Ashley, David H Liang, and James Y Zou. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In *NeurIPS ML4H Workshop: Vancouver, BC, Canada*, 2019. 5

[26] Nathan Painchaud, Nicolas Duchateau, Olivier Bernard, and Pierre-Marc Jodoin. Echocardiography segmentation with enforced temporal consistency. *IEEE Transactions on Medical Imaging*, 41(10):2867–2878, 2022. 2

[27] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197*, 2023. 2

[28] Hariharan Ravishankar, Rohan Patil, Vikram Melapudi, and Pavan Annangi. Sonosam-segment anything on ultrasound images. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 23–33. Springer, 2023. 2, 6, 7

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 6

[30] Ayesha Saadia and Adnan Rashdi. A speckle noise removal method. *Circuits, Systems, and Signal Processing*, 37:2639–2650, 2018. 1

[31] Xuemeng Song, Liqiang Jing, Dengtian Lin, Zhongzhou Zhao, Haiqing Chen, and Liqiang Nie. V2p: Vision-to-prompt based multi-modal product summary generation. In *Proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval*, pages 992–1001, 2022. 3

[32] Sarina Thomas, Andrew Gilbert, and Guy Ben-Yosef. Light-weight spatio-temporal graphs for segmentation and ejection fraction prediction in cardiac ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 380–390. Springer, 2022. 2

[33] Hongrong Wei, Heng Cao, Yiqin Cao, Yongjin Zhou, Wufeng Xue, Dong Ni, and Shuo Li. Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 623–632. Springer, 2020. 2

[34] Hongrong Wei, Heng Cao, Yiqin Cao, Yongjin Zhou, Wufeng Xue, Dong Ni, and Shuo Li. Temporal-consistent segmentation of echocardiography with co-learning from appearance and shape. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 623–632. Springer, 2020. 2

[35] Huisi Wu, Jiasheng Liu, Fangyan Xiao, Zhenkun Wen, Lan Cheng, and Jing Qin. Semi-supervised segmentation of echocardiography videos via noise-resilient spatiotemporal semantic calibration and fusion. *Medical Image Analysis*, 78:102397, 2022. 2

[36] Huisi Wu, Jingyin Lin, Wende Xie, and Jing Qin. Super-efficient echocardiography video segmentation via proxy- and kernel-based semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2803–2811, 2023. 2

[37] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 2, 6, 7

[38] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 2

[39] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 35:36324–36336, 2022. 2

[40] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. 2, 6, 7

[41] Rongzhi Zhang, Yue Yu, Pranav Shetty, Le Song, and Chao Zhang. Prboost: Prompt-based rule discovery and boosting for interactive weakly-supervised learning. *arXiv preprint arXiv:2203.09735*, 2022. 3

[42] Yichi Zhang and Rushi Jiao. How segment anything model (sam) boost medical image segmentation? *arXiv preprint arXiv:2305.03678*, 2023. 2