MemSAM: Taming Segment Anything Model for Echocardiography Video Segmentation

Supplementary Material

1. More Implementation Details

Ejection Fraction Estimation. Following the guidelines from the American Society of Echocardiography [1], left ventricular volume was estimated using Simpson's biplane method of disks. Simpson's method necessitates simultaneous estimation across apical two-chamber (a2c) and four-chamber (a4c) view videos. The ventricle is partitioned into multiple disks within each view. The volume of each disk is calculated individually and then summed together. The left ventricular volume V calculation is formally defined as:

$$V = \frac{\pi}{4} \sum_{1}^{n} D_i^{2c} \times D_i^{4c} \times \frac{L}{n} \tag{1}$$

where D_i^{2c} and D_i^{4c} denote the chamber diameters across the two-chamber and four-chamber apical views respectively, and L indicates the length of the long axis. n is the number of disks (typically set to 20). Finally, the calculation of left ventricular ejection fraction LV_{EF} is as follows:

$$LV_{EF} = \frac{V_{ED} - V_{ES}}{V_{ED}} \times 100\% \tag{2}$$

where V_{ED} and V_{ES} denote the volumes at the end-diastole (ED) and end-systole (ES) respectively.

Hyperparameter Settings. The loss $\mathcal{L}$ is calculated as follows:

$$\mathcal{L} = \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice} \tag{3}$$

where $\mathcal{L}_{bce}$ and $\mathcal{L}_{dice}$ are binary cross entropy loss and dice loss [2] respectively. We use $\lambda_{bce} = 0.2$ and $\lambda_{dice} = 0.8$. We use a polynomial decay learning rate adjustment strategy, where the learning rate is multiplied by $\left(1 - \frac{iter}{iter_{max}}\right)^{0.9}$ at each iteration. Our method is trained and evaluated with batch size 1 on a single RTX 3090 GPU. To ensure the repeatability of the experiment, we set all random number seeds to 1234.

Initialization Memory Strategy. To preclude erratic segmentation quality, we performed additional processing on the first frame. Point prompts are initially utilized by the SAM component to generate a coarse mask. This coarse mask is then refined into a refined mask through the memory prompt, and then the refined mask is added to the memory bank. Initializing memory using the refined mask maintains continuity between the first frame prediction and subsequent frame predictions.

2. More Experiments

Different Numbers of Labeled Frames. We conducted ablation studies on different numbers of labeled frames, as shown in Table 1. For the 10-frame videos, we experimented with different numbers (2,4,6,8,10) of annotated frames during training. The results demonstrate that additional labeled frames improve performance with our method, though gains significantly diminish after sparse annotations. Notably, the use of just two annotated frames achieves sufficiently accurate segmentation.

label	mDice ↑	mIoU ↑	HD95 ↓	ASSD ↓
2-frame	93.31	87.61	3.82	1.57
4-frame	93.45	87.88	3.74	1.53
6-frame	93.59	88.06	3.67	1.50
8-frame	93.70	88.23	3.61	1.48
10-frame	**93.79**	**88.46**	**3.58**	**1.45**

Table 1. Performance of different numbers of labeled frames on the CAMUS-Semi test set. "n-frame" denotes the utilization of n annotated frames during training.

Learning Rate. We also tried different base learning rates (baselr). The results are shown in the Table 2. When baselr = 1e-4, the segmentation performance is the best.

baselr	mDice ↑	mIoU ↑	HD95 ↓	ASSD ↓
1e-5	93.10	87.22	3.97	1.64
5e-5	93.28	87.59	3.85	1.60
1e-4	**93.31**	**87.61**	**3.82**	**1.57**
5e-4	92.28	85.88	4.15	1.88

Table 2. Performance of different base learning rates on the CAMUS-Semi test set.

3. More Visual Comparison Results

We demonstrate a qualitative comparison between our method and other SOTA methods on a full 10-frame video. As shown in Figure 1 and 2. With insufficient labeled data for fully supervised training, state-of-the-art approaches exhibit considerable erroneous regions across segmentation predictions due to a lack of guidance. In contrast, by leveraging memory prompt and reinforcement, our proposed framework demonstrates precise delineation of anatomical structures across almost all frames.

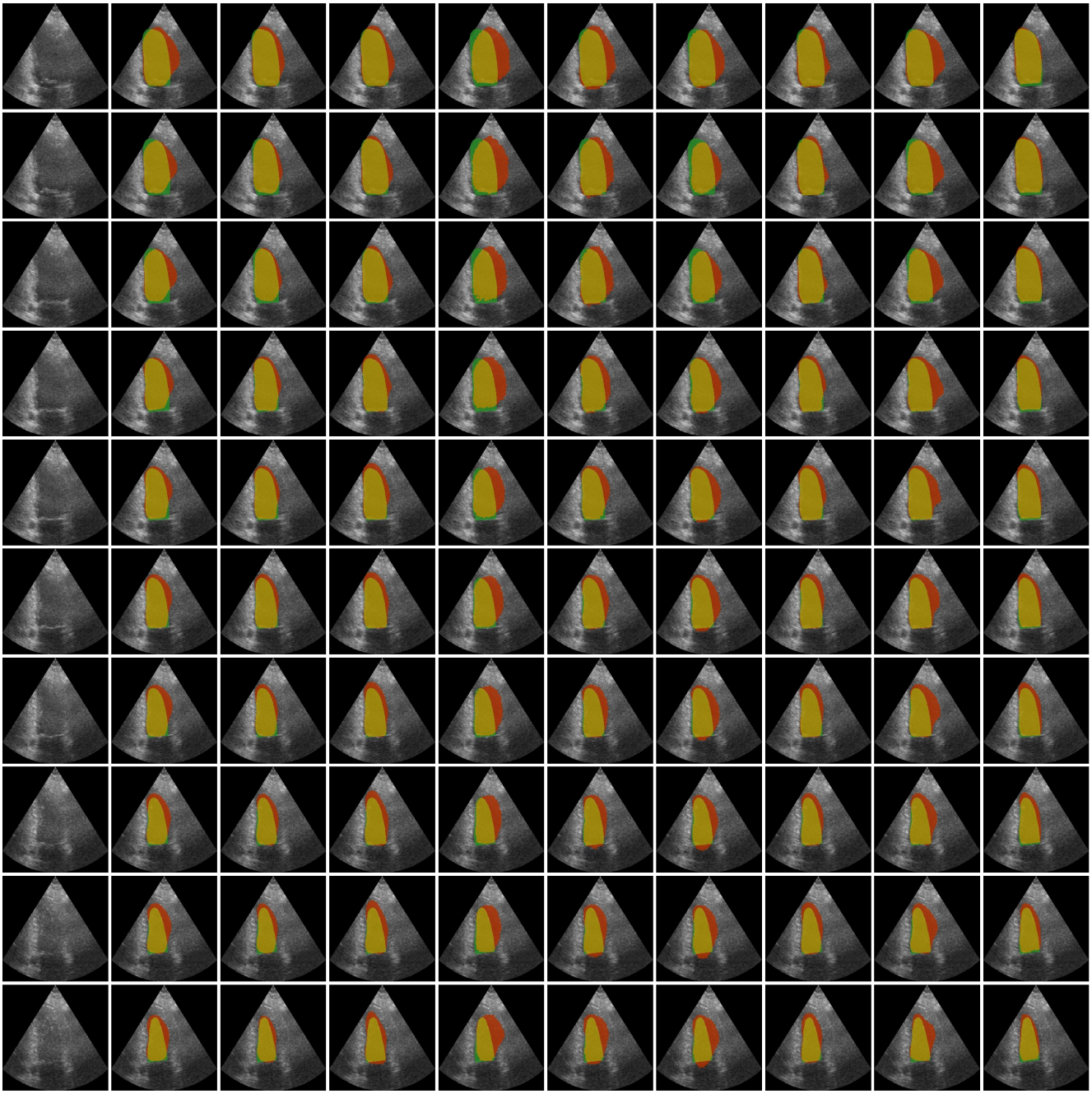

Figure 1. More visual comparison results of our method with other SOTA methods on the CAMUS-Semi test set. Each column shows the predictions of one method in chronological order. Green, red, and yellow regions represent the ground truth, prediction, and overlapping regions, respectively.

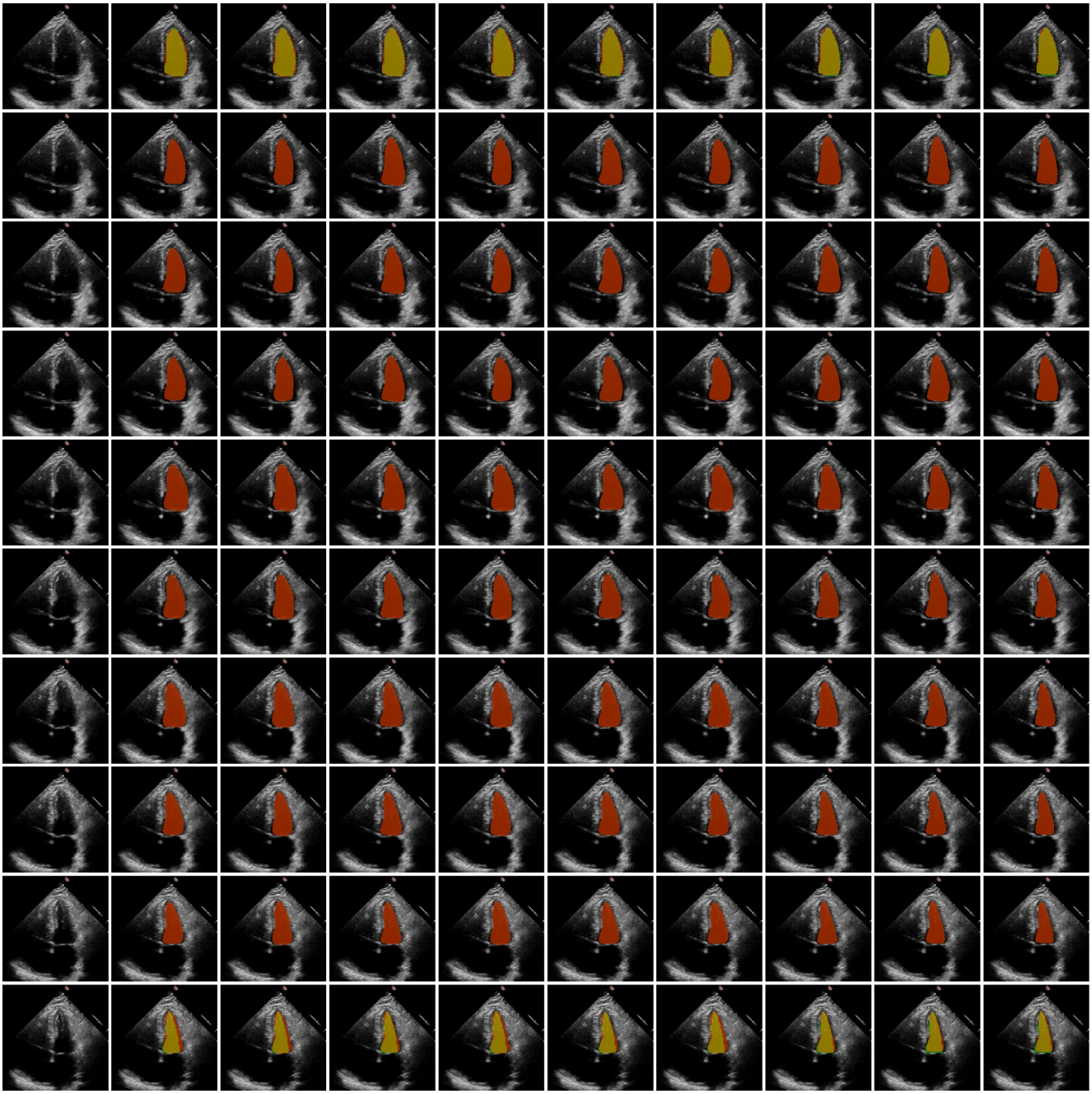

Figure 2. More visual comparison results of our method with other SOTA methods on the EchoNet-Dynamic test set. Each column shows the predictions of one method in chronological order. Green, red, and yellow regions represent the ground truth, prediction, and overlapping regions, respectively.

References

[1] Roberto M Lang, Luigi P Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A Flachskampf, Elyse Foster, Steven A Goldstein, Tatiana Kuznetsova, 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. *European Heart Journal-Cardiovascular Imaging*, 16(3):233–271, 2015. 1

[2] 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. 1