Supplementary Material for

Beyond Adapting SAM: Towards End-to-End Ultrasound Image Segmentation via Auto Prompting

Table 1: Summary of the datasets in US30K. LV, MYO, and LA are short for the left ventricle, myocardium, and left atrium.

Dataset	Slice number	Mask number	Train slice	Validation slice	Test slice	Segmentation target
TN3K	3493	3493	2303	576	614	Thyroid nodule
DDTI	637	637	-	-	637	Thyroid nodule
TG3K	3585	3585	3226	359	-	Thyroid gland
BUSI	647	647	454	64	129	Breast cancer
UDIAT	163	163	-	-	163	Breast cancer
CAMUS	19232	57696	15315	1949	1968	LV, MYO, LA
HMC-QU	2349	2349	-	-	2349	MYO
US30K	30106	68570	21298	2948	5860	Above 6 categories

Table 2: Quantitative comparison of our SAMUS and SOTA task-specific methods on segmenting thyroid nodule (TN3K), breast cancer (BUSI), left ventricle (CAMUS-LV), myocardium (CAMUS-MYO), and left atrium (CAMUS-LA). The performance is evaluated by the Dice score (%) and Hausdorff distance (HD). The best results are marked in bold.

Method	TN3K		BUSI		CAMUS-LV		CAMUS-MYO		CAMUS-LA	
	Dice	HD	Dice	HD	Dice	HD	Dice	HD	Dice	HD
U-Net	79.01	34.12	78.11	33.60	93.56	9.90	86.86	16.87	91.00	12.91
CPFNet	79.43	33.07	80.56	27.98	93.32	9.63	86.68	16.51	91.51	12.26
CA-Net	80.52	33.65	81.88	28.67	93.59	9.77	87.21	**16.24**	91.28	12.22
CE-Net	80.37	32.79	81.60	29.19	93.31	9.65	86.47	16.66	91.14	12.39
AAU-Net	82.28	30.53	80.81	28.96	93.32	9.97	86.98	16.49	91.35	12.12
SwinUnet	70.08	44.13	67.23	47.02	91.72	12.80	84.46	20.25	89.80	14.74
SETR	67.80	44.11	68.22	40.37	92.82	11.34	86.20	18.27	90.52	13.91
MISSFormer	79.42	32.85	78.43	33.10	93.25	9.94	86.57	16.50	91.18	11.82
TransUNet	81.44	30.98	82.22	27.54	93.54	9.60	87.20	16.36	91.37	12.10
TransFuse	78.50	32.44	73.52	34.95	93.30	10.07	86.77	17.25	90.68	12.46
FAT-Net	80.45	32.77	82.16	28.55	93.59	**9.20**	87.19	15.93	91.55	12.05
H2Former	82.48	30.58	81.48	27.84	93.44	9.79	87.31	16.60	90.98	11.92
SAMUS	**84.45**	**28.22**	**85.77**	**25.49**	**93.73**	9.79	**87.46**	16.74	**91.58**	**11.60**

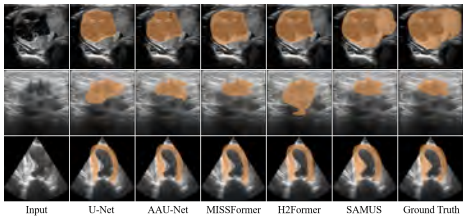

Fig. 1: Qualitative comparisons between SAMUS and task-specific methods. From top to bottom are examples of segmenting thyroid nodule, breast
and myocardium.

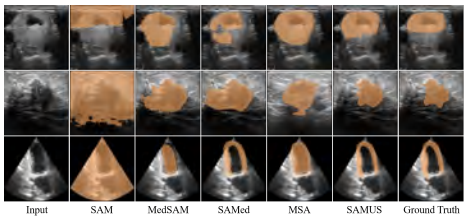

Fig. 2: Qualitative comparisons between SAMUS and foundation models. From top to bottom are examples of segment-

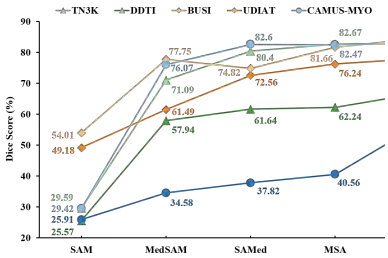

Fig. 3: Segmentation and generali
ability comparison of our SAMUS and other foundation models on seeable (in light color) and unseen (in dark color) US30K data.

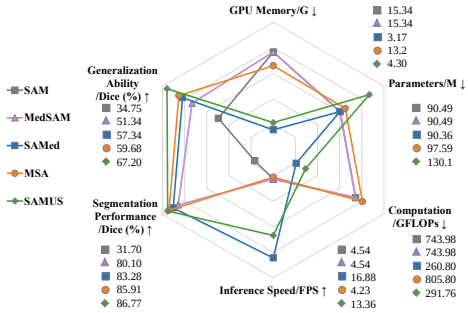

Fig. 4: Comparison of SAMUS and foundation models on GPU memory cost, model parameters, computational complexity, inference speed, performance, and generalization.

Table 3: Ablation study on the task token number k of APG on DDTI. The prompt embeddings generated by a small number of task tokens are not accurate enough, and when the number of task tokens reaches a certain level, performance tends to saturate.

k	Dice (%)	HD	IoU	ACC	SE	SP
1	77.63	31.06	66.15	96.45	82.74	97.89
10	**78.89**	**30.17**	**68.37**	**96.73**	82.42	**98.20**
20	78.49	30.83	67.62	96.45	**83.39**	97.74
50	78.70	30.74	67.58	96.64	81.58	98.15