Enhancing SAM with Efficient Prompting and Preference Optimization for Semi-supervised Medical Image Segmentation

Supplementary Material

1. More Qualitative Results

We present segmentation maps for 3 datasets in Fig. 7 and Fig. 6, generated by our model trained at 50% data settings. The saliency maps could correctly highlight the target regions and proved to be an excellent source of supervision. It can be noted from Fig. 7 that our segmentation quality around the boundaries of tumor or lung is much superior compared to both nnUnet and SAM-Med2D. In Fig. 6, similar trends can be seen while segmenting the abdominal organs – right kidney, bladder, and aorta (top to bottom).

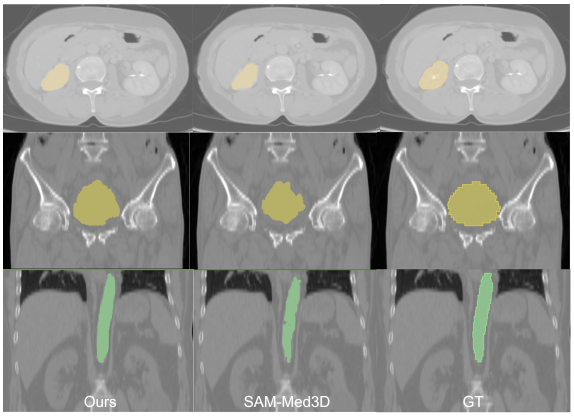

Figure 6. Segmentation maps of three anatomical structures (right kidney, bladder, and aorta)↓ for SAM-Med3D and our method

2. Robustness to noise in rating

We randomly flipped one of three rating combinations (1↔2, 2↔3,3↔4) for 5-30% of the training image samples. This was done to evaluate the robustness of our framework to noise in the rating process. Despite the introduction of noise through the virtual annotator, the Dice scores showed minimal decline. The results have been shown in Tab. 4. With 30% of the image samples affected by noisy ratings, the dice score performance decreased by only 0.24, 0.20, and 0.24 for the X-ray, USD, and CT datasets, respectively.

Flip (%)	Dice score (20% data)		
	Chest-Xray	Breast-USD	AMOS-CT (mean)
0	**78.87**	**75.88**	**77.69**
5	78.82	75.83	77.62
10	78.79	75.81	77.58
20	78.71	75.74	77.51
30	78.63	75.68	77.45

Table 4. Ablation results for varying proportions (5%-30%) of images with flipped ratings.

3. Prompt design

Text-based prompts were designed to provide inputs for the BiomedCLIP, MedVInT, and GPT-4 models, enabling both direct and indirect supervision from them. This supervision can take the form of responses or guidance for generating saliency maps. A summary of the design for each of the three datasets is provided in the Tab. 3.

4. Quantitative comparison with SOTA

The quantitative results corresponding to the line graphs in Fig. 3 are reported in Tables 5, 6, and 7 respectively. The results indicate that our proposed method, owing to unsupervised prompting and human-in-the-loop feedback, outperforms SOTA approaches in low-data regimes, demonstrating superior generalization and robustness.

5. Example answers from the VQA branch

The following VQA outputs illustrate the model's ability to accurately locate target objects based on text prompts:

- The left apical and anterior lobes of the lung are affected by COVID infiltrates in the image.

VLM	**Prompts**		
	Chest X-ray	**Breast USD**	**AMOS-CT**
BiomedCLIP	chest x-ray	[class] breast tumor	[organ]
MedVInT	Briefly describe the condition of lungs and location of pathologies	What is the shape of breast tumor and where is it located?	What is the shape of the [organ] and where is it located?
GPT-4	Briefly describe, in one line, the lungs of a patient suffering from [disease]	Briefly describe, in one line, [class] breast tumor of a patient in Ultrasound	Briefly describe, in one line, [organ] of a human in CT

Table 3. Different prompts designed for the BiomedCLIP, MedVInT, and GPT-4 models. The placeholder [class] refers to the tumor type, either malignant or benign, while [organ] refers to one of the 15 organs available in the AMOS-CT dataset for segmentation.

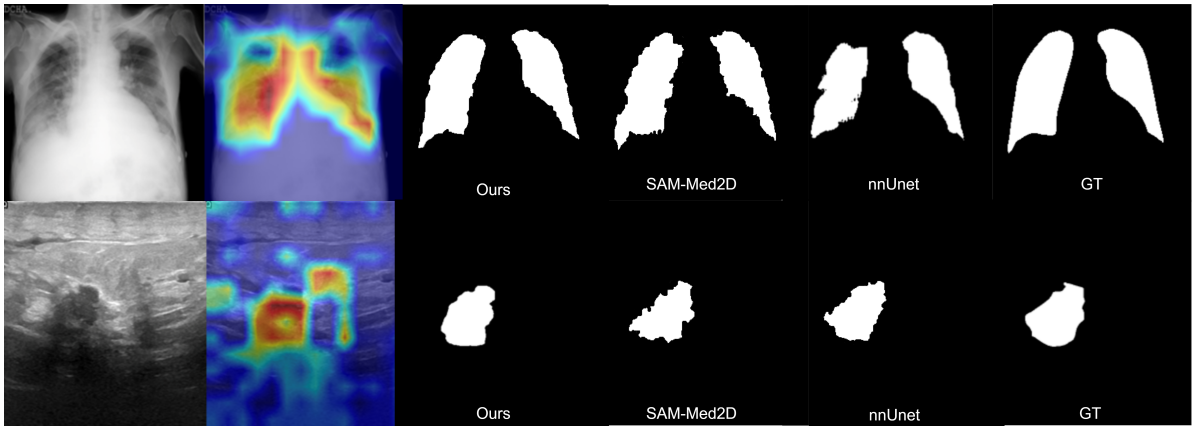

Figure 7. Qualitative comparisons were made between the segmentation results of nnUnet, SAM-Med2D, and our framework on 2D datasets. BiomedCLIP-based saliency maps are also depicted. Experiments were conducted in 50% data settings.

Methods	Dice (CXR)			
	10%	20%	50%	100%
Unet	47.64	58.66	74.19	95.82
nnUnet	48.72	60.97	75.50	96.38
SAM	49.81	61.64	74.30	96.03
SAMMed2D	54.25	67.81	79.49	96.72
Self-prompt	56.39	68.41	81.56	96.20
Ours (Prompt)	75.60	79.13	91.42	96.27
Ours (Prompt + Feedback)	75.60	78.87	89.68	95.27

Table 5. Quantitative comparison with SOTA. Dice score for Chest Xray on different proportions of training data.

Methods	Dice (USD)			
	10%	20%	50%	100%
Unet	38.62	49.16	70.24	93.06
nnUnet	41.30	53.57	75.41	94.15
SAM	43.97	66.45	78.84	93.59
SAMMed2D	52.13	75.64	86.97	94.20
Self-prompt	54.90	76.42	87.06	93.65
Ours (Prompt)	73.62	81.38	89.04	95.73
Ours (Prompt + Feedback)	73.62	75.88	88.15	94.07

Table 6. Quantitative comparison with SOTA. Dice score for Breast USD on different proportions of training data.

- The ultrasound identifies a 1.8 cm $\times$ 1.5 cm irregular hypoechoic mass located in the upper outer quadrant of the image.
- The liver is of normal size, measuring approximately 15 cm in craniocaudal dimension. It is located in the right upper quadrant of the abdomen, extending from the diaphragm down to the upper poles of the kidneys.

Methods	mDice (AMOS)			
	10%	20%	50%	100%
Unet	54.29	59.35	67.92	76.54
nnUnet	58.72	65.21	78.03	88.88
SAM	57.20	64.93	71.28	80.67
SAMMed2D	62.48	66.57	77.36	88.14
SAMMed3D	67.54	72.54	83.27	89.56
Self-prompt	63.72	66.83	78.29	86.17
Ours (Prompt)	74.77	79.20	86.36	91.04
Ours (Prompt + Feedback)	74.77	77.69	85.70	89.80

Table 7. Quantitative comparison with SOTA. Mean dice score for AMOS CT on different proportions of training data.

6. Additional dataset and implementation details

For **Ultrasound Breast Tumor Segmentation**, the dataset consists of 810 images, with 437 benign and 210 malignant cases from BUSI [3] and 109 benign and 54 malignant cases from UDIAT [9]. Of these, 600 images were used for training and 210 for testing.

For **Chest X-ray Lung Segmentation**, 27,132 images from COVID-QU-Ex [12] were used, with a separate test set of 6,788 images. The number of image samples per disease label for train and test sets are: COVID (9,561, 2,395), non-COVID (9,010, 2,253), and Normal (8,561, 2,140).

For **Abdominal CT Organ Segmentation**, the model was trained on 200 annotated CT scans from the AMOS-CT training set and evaluated on 100 CT scans from the validation set.

While incorporating the preference alignment module, we use the Intersection over Union (IoU) scores between the predicted masks and the ground truth to generate ratings and/or rankings. The IoU scores are binned into the follow-

ing ranges: {<0.4, 0.4-0.55, 0.55-0.7, and >0.7}. For the generation of multiple segmentation proposals, the model's output probabilities are thresholded at 0.3, 0.4, 0.5, and 0.6. The loss function for training this second stage is listed in Eqn. 5. The weights $\beta_1 = 1$ and $\beta_2 = 0.5$ were experimentally determined to be optimal. We tested different pairs of β_1, and β_2 values to identify the optimal combination in Eqn. 5. As shown in Tab. 8, the best performance was achieved with $\beta_1 = 1$, and $\beta_2 = 0.5$.

β_1	β_2	Dice score (20% data)		
		Chest-Xray	Breast-USD	AMOS-CT (mean)
2	1	78.12	75.43	77.47
1.5	0.75	78.64	75.70	77.53
1	**0.5**	**78.87**	**75.88**	**77.69**

Table 8. Selection of experimental parameters β_1, β_2