PRISM: A Promptable and Robust Interactive Segmentation Model with Visual Prompts

Hao Li, Han Liu, Dewei Hu, Jiacheng Wang, and Ipek Oguz

Vanderbilt University

Abstract. In this paper, we present PRISM, a **P**romptable and **R**obust **I**nteractive **S**egmentation **M**odel, aiming for precise segmentation of 3D medical images. PRISM accepts various visual inputs, including points, boxes, and scribbles as sparse prompts, as well as masks as dense prompts. Specifically, PRISM is designed with four principles to achieve robustness: (1) Iterative learning. The model produces segmentations by using visual prompts from previous iterations to achieve progressive improvement. (2) Confidence learning. PRISM employs multiple segmentation heads per input image, each generating a continuous map and a confidence score to optimize predictions. (3) Corrective learning. Following each segmentation iteration, PRISM employs a shallow corrective refinement network to reassign mislabeled voxels. (4) Hybrid design. PRISM integrates hybrid encoders to better capture both the local and global information. Comprehensive validation of PRISM is conducted using four public datasets for tumor segmentation in the colon, pancreas, liver, and kidney, highlighting challenges caused by anatomical variations and ambiguous boundaries in accurate tumor identification. Compared to state-of-the-art methods, both with and without prompt engineering, PRISM significantly improves performance, achieving results that are close to human levels. The code is publicly available at https://github.com/MedICL-VU/PRISM.

Keywords: Interactive medical image segmentation · Iterative correction · Tumor segmentation · Visual prompts · Segment Anything Model (SAM).

1 Introduction

In recent years, deep learning-based methods have achieved state-of-the-art performance in various segmentation tasks [21,12]. However, achieving robust outcomes remains a challenge due to significant anatomical variations among individuals and ambiguous boundaries in medical images. Alternatively, interactive segmentation models, such as the Segment Anything Model (SAM) [14], offer a solution by involving humans in the loop and utilizing their expertise to achieve precise segmentation. This requires users to indicate the target region by providing visual prompts, such as points [22,14,23,9,18,17,30,8], boxes [32,35,7,29,14,24,3,34], scribbles [30,4,33,6], and masks [14,27,31,5,22].

Importantly, a robust interactive segmentation model should be effective in responding to visual prompts given by users with minimal interactions. Inspired by SAM [14], many interactive segmentation methods have been proposed in medical imaging. However, most of these SAM-based methods are limited to using a single type of prompt [32,9,34,8,18,7,24,3,35]. This limits the available information, impacting the effectiveness of the model. Due to the heterogeneity in anatomy and appearance of medical images, the models should leverage the advantages of different prompts to achieve optimal performance across various medical applications. Moreover, these interactive methods do not involve humans in the loop [9,8,35,18,7,32,24,3,34], i.e., the visual prompts are applied to the automatic segmentation model only once, which may not be sufficient to achieve robust outcomes with the initial prompts provided. In contrast, practical applications often necessitate iterative corrections based on new prompts from the user until the outcome meets their criteria [26]. Unfortunately, few studies have explored such human-in-loop approach for medical interactive segmentation [5,31], and their performance has been suboptimal.

An additional concern is the efficiency of the model with respect to training and user interaction requirements. Training a model from scratch with extensive datasets [5,33,24,31] improves overall representational capabilities, but this is time-consuming and typically requires substantial computational resources for optimal performance. Furthermore, models may lack specificity for particular applications due to existing domain gaps in medical imaging [31]. This requires an increased number of user prompts to reach adequate performance. Moreover, 2D models [5,33,24] are not considered "efficient" for 3D images as they require significant user effort to provide visual prompts for each slice.

In this work, we propose a robust method for interactive segmentation in medical imaging. We strive for human-level performance, as a human-in-loop interactive segmentation model with prompts should gradually refine its outcomes until they closely match inter-rater variability [10]. We present PRISM, a **P**romptable and **R**obust **I**nteractive **S**egmentation **M**odel for 3D medical images, which accepts both sparse (points, boxes, scribbles) and dense (masks) visual prompts. We leverage an iterative learning [14,27] and sampling strategy [26] to train PRISM to achieve continual improvements across iterations. The sparse visual prompts are sampled based on the erroneous region of the prediction from previous iteration, which simulates the human correcting behavior. For each iteration, multiple segmentation masks are generated [19] along with regressed confidence scores. The output with the highest score is selected, increasing the robustness of the model. This approach is similar to model ensembling [20]. Moreover, inspired from [25,13], the output is fed into a shallow corrective refinement network to correct the mislabeled voxels and refine the final segmentation, adhering to the goal of being effectiveness. A hybrid image encoder with parallel convolutional and transformer paths [16] is used as the backbone to better capture the local and global information from a medical image. Our main contributions are summarized as:

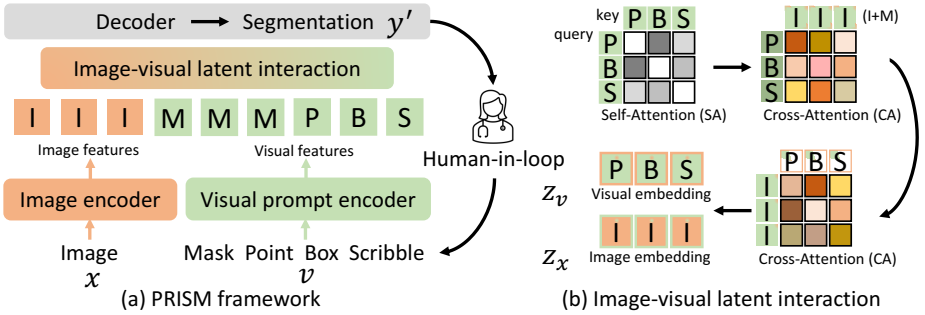

Fig. 1: **(a)** PRISM takes an image (x) and visual prompts (v) to produce a segmentation (y'). The user then provides prompts for next iteration. **(b)** Interaction between image and visual features in the latent space to produce image (Z_x) and visual (Z_v) embeddings with self- and cross-attention mechanisms.

- Interaction: PRISM accepts various visual prompts, offering versatility that effectively addresses the diverse challenges of medical imaging segmentation. This approach ensures precise outcomes through user-friendly interaction.
- Human-in-loop: PRISM is trained with iterative and confidence learning, allowing for continuous improvements and robust performance. With each user input, the performance progressively approaches expert-level precision.
- Network architecture: PRISM employs a corrective refinement network and a hybrid encoder to produce precise segmentation for challenging conditions.

We evaluate PRISM on four public tumor datasets, including tumors in colon [1], pancreas [1], liver [2] and kidney [11], where anatomical differences among individuals and ambiguous boundaries are present. Comprehensive validation is performed against state-of-the-art automatic and interactive methods, and PRISM significantly outperforms all of them with substantial improvements.

2 Methods

PRISM employs a generic encoder-decoder architecture as displayed in Fig. 1(a). For a given input image x, along with visual prompts v, the image and prompt encoders generate the image and visual features, respectively. These latent features then interact through self- and cross-attention mechanisms (Fig. 1(b)). The resulting embeddings, Z_x and Z_v, are fed into the decoder to produce the final segmentation y'. The expert then gives sparse prompt based on the erroneous regions in y' for next iteration, in human-in-loop manner.

Fig. 2(a) details PRISM, where a hybrid encoder combines parallel CNN and vision transformer paths to better extract image features. The last feature map of decoder f_d is integrated with the visual embedding to generate multiple mask predictions and associated confidence scores (Fig. 2(b)). A selector is employed to pick the candidate prediction with the highest confidence score, which is

subsequently fed into the corrective refinement network to generate the y'. In our experiments, we use a sampler to mimic expert corrective actions by identifying point and scribble prompts from the false positive and negative areas within y'.

Iterative learning. PRISM is designed for clinical applications, which require improvement through successive iterations. We define the loss for a single iteration as L_i, and the total loss as $L_{total} = \sum_{i=1}^{N} L_i$, where N denotes the total number of iterations. The input visual prompts come only from the current iteration, i.e., they are not cumulative. Since the dense prompt y'_i is generated from the previous iteration and retains the gradient, PRISM can learn the relationship between iterations. For efficiency, the image features are generated only once at the initial iteration. The following prompt types are allowed (Fig. 2(b)):

- At each iteration i, point prompts are randomly sampled with uniform distribution from the false negative (FN) and positive (FP) regions of y'_{i-1}.
- The scribbles are generated similar to Scribbleprompt [33] by first extracting the skeleton of the FN and FP regions of y'_{i-1} from the previous iteration. Next, a random mask is created by pixels randomly sampled from a normal distribution, applying Gaussian blur, and thresholding the image based on its mean value. This random mask is applied to divide the skeleton of y'_{i-1} into separate, smaller parts. Finally, a random deformation field and random Gaussian filtering are used to change the scribble curvature and thickness.
- The 3D bounding box (BB) is determined based on the ground truth y at the initial iteration ($i = 1$) and remains unchanged for the subsequent iterations.
- The logits map at iteration $i - 1$ is used to provide additional information instead of a binary mask as the dense prompt for iteration i, for $i > 1$.

Confidence learning. Unlike traditional methods that output a single mask, PRISM increases robustness by generating multiple outputs, each with a confidence score. The last feature map of the decoder, denoted as $f_d = Decoder(Z_x)$, interacts with the visual embedding to produce continuous maps m_j and confidence scores s_j via $m_j = f_d \times MLP_j^m(Z_v)$ and $s_j = MLP_j^s(Z_v)$, where MLP indicates the multi-layer perceptrons, and j is an index over the multiple outputs. Thus, m_j interacts with visual prompts at both latent- and image-level. The process of confidence supervision for each iteration is defined by: $L_{con} = \sum_{j=1}^{M}(\lambda_s L_s(m_j, y) + \lambda_b L_b(m_j, y) + \lambda_r L_r(s_j, 1 - L_{Dice}(m_j, y)))$, where L_s, L_b, L_r and L_{Dice} indicate structural, boundary, regression and Dice loss, respectively. More precisely, the structural information is captured through a combination of Dice loss and cross-entropy loss: $L_s = L_{Dice} + L_{CE}$. For a given logits map or binary mask, the smooth boundary is derived as: $B(m) = |m - P_{ave}(m)|$, where P_{ave} denotes the average pooling layer. The boundary loss (L_b) is evaluated by $L_{MSE}(B(m), B(y))$, with L_{MSE} representing the mean square error loss. We also use L_{MSE} for the regression loss.

Corrective learning. As shown in Fig. 2(a), a selector is employed to choose the candidate mask $\hat{m}$ corresponding to the highest confidence score s_j, which is then fed into the corrective refinement network. We design this network to be both shallow and efficient, consisting of two cascaded residual blocks and

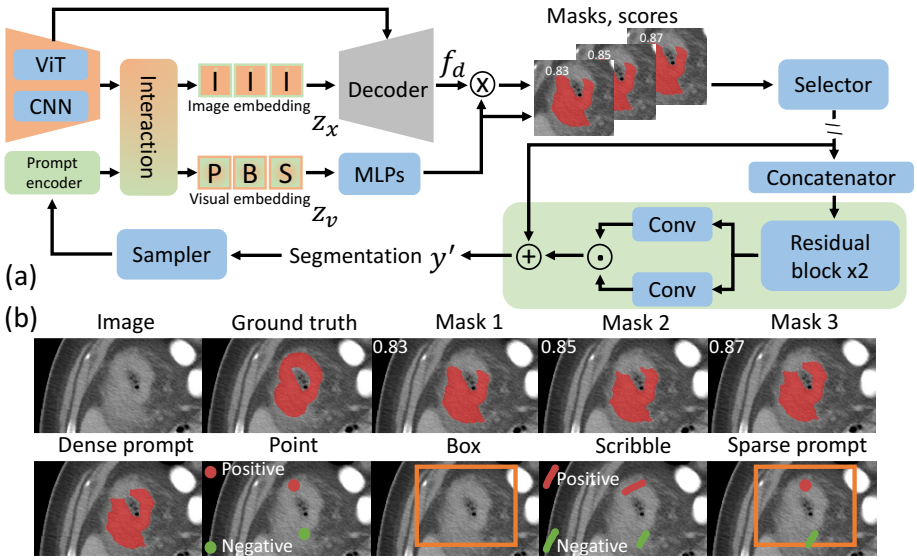

Fig. 2: **(a)** Details of PRISM. Green highlights the corrective refinement network. **(b)** Top row shows the multi-mask prediction with labeled confidence scores. The selector would then pick Mask 3 as the dense prompt. Possible visual prompts given this dense prompt are shown in the bottom row.

two convolution operations in two different branches. It takes the input x_c at original image resolution. Importantly, x_c is the concatenation of 4 images: (1) the input image x, (2) the binary mask $\hat{b}$ derived from $\hat{m}$, (3) the binary positive prompt maps accumulated from iteration 0 to i, and (4) the similar binary cumulative negative prompt map. For efficiency, the x_c is downsampled and then upsampled back for y'_i. The final segmentation y'_i is supervised by: $L_{cor} = \lambda_s L_s(y'_i, y) + \lambda_b L_b(y'_i, y)$. Notably, the corrective network does not backpropagate to the network weights for candidate mask generation and focuses exclusively on refining any given logit map. The y'_i is fed to sampler for next iteration $i+1$.

Implementation details. For training, the total objective function is: $L_{total} = \sum_{i=1}^{N}(L_{con_i} + L_{cor_i})$, where the ratio of $\lambda_s : \lambda_b : \lambda_r$ as $1 : 10 : 1$. The batch size was set to 2, and the initial learning rate was $4e^{-5}$, which was reduced by $2e^{-6}$ after every epoch. We employed the AdamW optimizer across a maximum of 200 epochs. Our study was conducted on an NVIDIA A6000. The input image was cropped to a 3D $128 \times 128 \times 128$ patch centered on a foreground pixel. The number of iterations (N) was fixed at 11, with 1 to 50 point prompts randomly chosen per iteration during training. The Dice and normalized surface Dice (NSD) are used as evaluation metrics. We compared to both fully automated [12,28,15] and interactive [9,18,31,14] methods. We used $y'_{i=11}$ as final segmentation for the iterative methods [31] and PRISM. We retrained using the published code of each model, and the pretrained weights were also employed if publicly available.

Table 1: Quantitative results. Bold indicates best performance. PRISM-plain is a simplified model for a fair comparison to methods not allowing BB prompts [9,18,31], and methods[14,9,18,31] only using 1 point prompt per input. For liver tumor, 10 points are used for all methods to accommodate multiple tumors.

	Public datastes (Dice % / NSD %)			
Methods	Colon tumor	Pancreas tumor	Liver tumor	Kidney tumor
nnU-Net [12]	43.91 / 52.52	41.65 / 62.54	60.10 / 75.41	73.07 / 77.47
3D UX-Net [15]	28.50 / 32.73	34.83 / 52.56	45.54 / 60.67	57.59 / 58.55
Swin-UNETR [28]	35.21 / 42.94	40.57 / 60.05	50.26 / 64.32	65.54 / 72.04
SAM [14]	28.83 / 33.63	24.01 / 26.74	8.56 / 5.97	36.30 / 29.86
3DSAM-adapter [9]	57.32 / 73.65	54.41 / 77.88	56.61 / 69.52	73.78 / 83.86
ProMISe [18]	66.81 / 81.24	57.46 / 79.76	58.78 / 71.52	75.70 / 80.08
SAM-Med3D [31]	54.34 / 78.58	65.61 / 92.40	23.64 / 26.97	76.50 / 88.41
SAM-Med3D-organ	70.75 / 91.03	76.40 / 97.75	66.52 / 77.97	88.20 / 97.80
SAM-Med3D-turbo	73.77 / 94.95	74.87 / 96.43	69.36 / 81.70	89.26 / 98.40
PRISM-plain	67.18 / 85.28	65.73 / 89.51	79.70 / 91.60	85.29 / 93.55
PRISM-ultra	**93.79 / 99.96**	**94.48 / 99.99**	**94.18 / 99.99**	**96.58 / 99.80**

"PRISM-plain" only uses 1 point and no BB. "-ultra" adds the BB, and scribbles.

We present two versions of our method. PRISM-plain only uses point prompts, while PRISM-ultra can handle other sparse visual prompts such as boxes and scribbles. All results are generated with seeds.

3 Experiments

Datasets. We use four public 3D CT datasets, including colon (N=126) [1], pancreas (N=281) [1], liver (N=118) [2] and kidney (N=300) [11] tumors. Challenges include large anatomical differences among individuals, varying shapes of target structures, and ambiguous boundaries. Additionally, multiple tumors [11,2] may occur in a single subject. We adopted the same data split as a prior study [9] for each task with a training/validation/testing split of 0.7/0.1/0.2, and we only use tumor labels to focus on binary segmentation. We resampled to a $1mm$ isotropic resolution, performed intensity clipping based on the 0.5 and 99.5 percentiles of the foreground, and applied Z-score normalization based on foreground voxels. Random zoom and intensity shift are used as data augmentations.

Comparison with state-of-the-art. Table 1 compares quantitative segmentation results (See qualitative results in Fig. S.1). As expected, fully automated methods [12,28,15] are unable to deliver satisfactory outcomes for these challenging datasets. Even with visual prompts, SAM [14] struggles with the substantial domain shift between natural and medical images. Furthermore, the interactive segmentation models [9,18,31] trained from scratch fail to produce adequate results. Notably, "-organ" and "-turbo" denote the fine-tuning of the pretrained

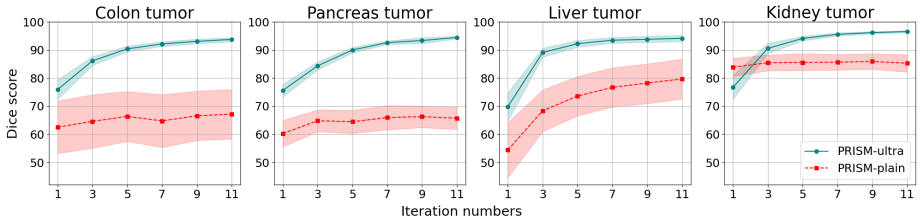

Fig. 3: Dice score of proposed PRISM on four tumor datasets, where the mean values (lines) and their 95% confidence intervals (shades) are presented.

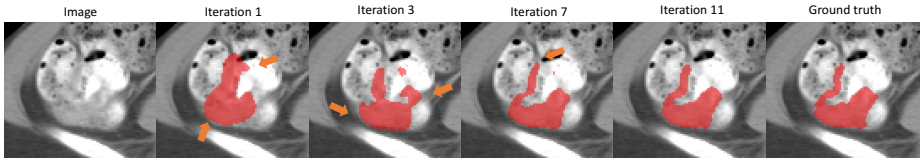

Fig. 4: Qualitative results of PRISM-ultra for colon tumors characterized by irregular shapes and ambiguous boundaries. The orange arrows indicate the major defects which are corrected in the subsequent iteration. The initial output has noticeable errors that rapidly get corrected in the first few iterations. More qualitative results can be viewed in Fig. S.2.

SAM-Med3D model [31]. "-organ" focuses on organ-specific datasets, while "-turbo" includes these organ datasets (including, in some cases, the test data in their pretraining) alongside a broader range of other medical imaging data. In contrast, excluding the pretrained models, the proposed PRISM-plain demonstrates superior overall Dice performance with only a single point as sparse prompt. Moreover, as more visual prompts (points, BB, and scribbles) are provided, PRISM-ultra delivers outcomes close to human level performance, demonstrating its strength as an effective human-in-loop algorithm.

Analysis of iterative learning performance. Fig. 3 shows the performance of PRISM across the iterative learning process. Although PRISM-plain delivers superior outcomes (Table 1), it does not have monotonically increasing performance across iterations and can suffer from a decrease in performance at later iterations, except for liver tumor. In practice this would be frustrating to the human user, as more input prompts paradoxically lowers the segmentation accuracy. However, PRISM-ultra presents not just substantial improvements in overall performance, but also a monotonically increasing accuracy and narrower 95% confidence intervals. This highlights the PRISM-ultra as a robust model, and its qualitative results in Fig. 4 and Fig. S.2 illustrate the iterative correction. Initially, the results may not meet expectations due to ambiguous boundaries and anatomical variations. Yet, as iterations advance, the outcomes progressively align more closely with the human label. Major corrections occur at the early stages which indicate the efficiency and effectiveness of PRISM-ultra.

Table 2: Ablation study on the colon tumor. ConL: confidence learning. PRISM-plain-b: PRISM-plain plus BB. Model configurations are detailed in Table S.1.

	ViT encoder	CNN encoder	Hybrid	PRISM-plain	PRISM-plain-b	no ConL
Dice	58.38	63.58	66.40	67.18	76.40	73.65

Table 3: Prompt analysis for colon tumors. "-1", "-50": number of points used during training. PRISM-basic: PRISM-plain-b with random number of points. "-erode" and "-dilate": modified BB by 5 voxel radius. CorL: corrective learning. PRISM-ultra: PRISM-basic plus test-time scribbles. PRISM-ultra+: PRISM-ultra plus train-time scribbles. Prompt configurations are detailed in Table S.2.

Test Points	-plain-b-1	-plain-b-50	-basic	no CorL	-erode	-dilate	-ultra	-ultra+
1	76.40	61.33	76.66	77.65	72.99	75.47	93.79	96.47
10	77.71	82.42	83.85	80.12	82.85	83.72	93.95	96.54
50	72.52	89.91	89.36	81.60	89.99	89.42	94.07	96.73
100	69.60	91.30	90.26	81.20	91.06	90.70	94.28	96.93

Analysis of framework. An ablation study for various frameworks on the colon dataset are presented in Table 2, and the details of model configurations in Table S.1. The CNN encoder outperforms the use of a Vision Transformer (ViT) alone. Using a hybrid encoder improves results, while proposed confidence and corrective learning further increase PRISM-plain performance. PRISM-plain is advanced to PRISM-plain-b through the incorporation of a 3D BB, a favored prompt in medical segmentation due to its strong image prior, leading to improved outcomes. For efficiency, drawing a box does not require significantly more effort than point prompt. For effectiveness, points can address the limited flexibility of boxes in making corrections in subsequent iterations. The proposed confidence learning approach also contributes to the performance of the model.

Analysis of number of prompts. We analyze the performance with respect to the number of point prompts in Table 3. The compared model configurations are detailed in Table S.2. We first evaluate the number of prompts used for training. The model plain-b-1 uses 1 prompt for training, plain-b-50 uses 50, and basic uses a random number between 1 and 50. Too few or too many training points lowers performance, while the random selection is the most robust. We also note that corrective learning has a more pronounced contribution when more prompts are used. We next test the sensitivity to BB precision by dilating and eroding the true BB. Either eroding or dilating BB decreases performance compared to PRISM-basic when few points are used, but using more points resolves this issue. Adding test-time scribbles (ultra) and train-time scribbles (ultra+) both improves the performance of PRISM-basic. However, this improvement increases training time, making scribbles less practical for large datasets.

4 Conclusion

In this study, we propose PRISM, a **P**romptable **R**obust **I**nteractive **S**egmentation **M**odel for 3D medical images. It supports various visual prompts and applies diverse learning strategies to achieve performance comparable to that of human raters. This is demonstrated across four challenging tumor segmentation tasks, where existing state-of-the-art automatic and interactive segmentation approaches do not meet the expected standards. Future work will include adaptation of pretrained models, and learning from datasets with sparse annotations.
Acknowledgments. This work was supported, in part, by NIH U01-NS106845, and NSF grant 2220401.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
3. Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. arXiv preprint arXiv:2309.08842 (2023)
4. Chen, X., Cheung, Y.S.J., Lim, S.N., Zhao, H.: Scribbleseg: Scribble-based interactive image segmentation. arXiv preprint arXiv:2303.11320 (2023)
5. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
6. Cho, S., Jang, H., Tan, J.W., Jeong, W.K.: Deepscribble: interactive pathology image segmentation using deep neural networks with scribbles. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 761–765. IEEE (2021)
7. Deng, G., Zou, K., Ren, K., Wang, M., Yuan, X., Ying, S., Fu, H.: Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 368–377. Springer (2023)
8. Gao, Y., Xia, W., Hu, D., Gao, X.: Desam: Decoupling segment anything model for generalizable medical image segmentation. arXiv preprint arXiv:2306.00499 (2023)
9. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. arXiv preprint arXiv:2306.13465 (2023)
10. Haarburger, C., Müller-Franzes, G., Weninger, L., Kuhl, C., Truhn, D., Merhof, D.: Radiomics feature reproducibility under inter-rater variability in segmentations of ct images. Scientific reports **10**(1), 12688 (2020)
11. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical image analysis **67**, 101821 (2021)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

13. Jang, W.D., Kim, C.S.: Interactive image segmentation via backpropagating refinement scheme. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5297–5306 (2019)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
15. Lee, H.H., Bao, S., Huo, Y., Landman, B.A.: 3d UX-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. In: The Eleventh International Conference on Learning Representations (2023)
16. Li, H., Hu, D., Liu, H., Wang, J., Oguz, I.: Cats: Complementary cnn and transformer encoders for segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2022)
17. Li, H., Liu, H., Hu, D., Wang, J., Oguz, I.: Assessing test-time variability for interactive 3d medical image segmentation with diverse point prompts. arXiv preprint arXiv:2311.07806 (2023)
18. Li, H., Liu, H., Hu, D., Wang, J., Oguz, I.: Promise: Prompt-driven 3d medical image segmentation using pretrained image foundation models. arXiv preprint arXiv:2310.19721 (2023)
19. Li, Z., Chen, Q., Koltun, V.: Interactive image segmentation with latent diversity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 577–585 (2018)
20. Linmans, J., Elfwing, S., van der Laak, J., Litjens, G.: Predictive uncertainty estimation for out-of-distribution detection in digital pathology. Medical Image Analysis **83**, 102655 (2023)
21. Liu, H., Hu, D., Li, H., Oguz, I.: Medical image segmentation using deep learning. Machine Learning for Brain Disorders pp. 391–434 (2023)
22. Liu, Q., Xu, Z., Bertasius, G., Niethammer, M.: Simpleclick: Interactive image segmentation with simple vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22290–22300 (2023)
23. Luo, X., Wang, G., Song, T., Zhang, J., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S.: Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. Medical image analysis (2021)
24. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
25. Oguz, B.U., Shinohara, R.T., Yushkevich, P.A., Oguz, I.: Gradient boosted trees for corrective learning. In: International Workshop on Machine Learning in Medical Imaging (2017)
26. Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation (2021)
27. Sun, S., Xian, M., Xu, F., Yao, T., Capriotti, L.: Cfr-icl: Cascade-forward refinement with iterative click loss for interactive image segmentation. arXiv preprint arXiv:2303.05620 (2023)
28. Tang, Y., Yang, D., Li, W., Roth, H., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis (2022)
29. Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al.: Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE transactions on medical imaging **37**(7), 1562–1573 (2018)

30. Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al.: Deepigeos: a deep interactive geodesic framework for medical image segmentation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1559–1572 (2018)
31. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d. arXiv preprint arXiv:2310.15161 (2023)
32. Wei, X., Cao, J., Jin, Y., Lu, M., Wang, G., Zhang, S.: I-medsam: Implicit medical image segmentation with segment anything. arXiv preprint (2023)
33. Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V.: Scribbleprompt: Fast and flexible interactive segmentation for any medical image. arXiv preprint (2023)
34. Yao, X., Liu, H., Hu, D., Lu, D., Lou, A., Li, H., Deng, R., Arenas, G., Oguz, B., Schwartz, N., et al.: False negative/positive control for sam on noisy medical images. arXiv preprint (2023)
35. Zhang, Y., Hu, S., Jiang, C., Cheng, Y., Qi, Y.: Segment anything model with uncertainty rectification for auto-prompting medical image segmentation. arXiv preprint arXiv:2311.10529 (2023)

Supplementary Materials

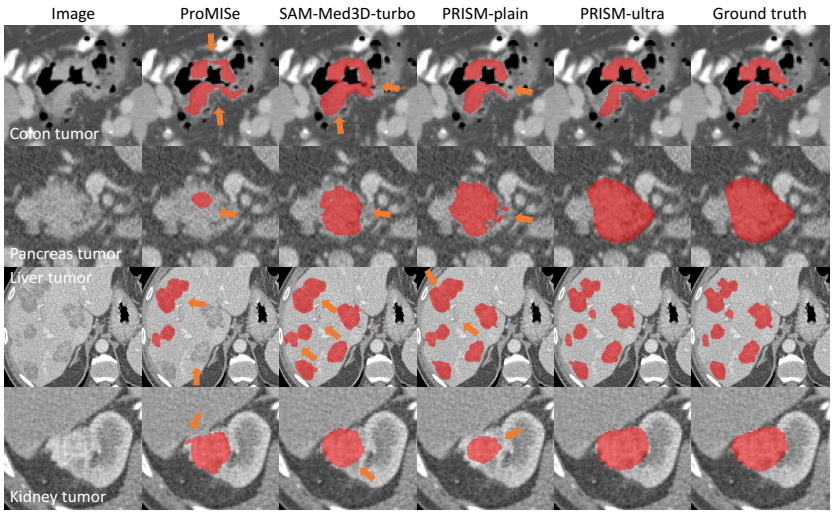

Fig. S.1: Qualitative results of four different tumor segmentation tasks. The orange arrows indicate the major defects.

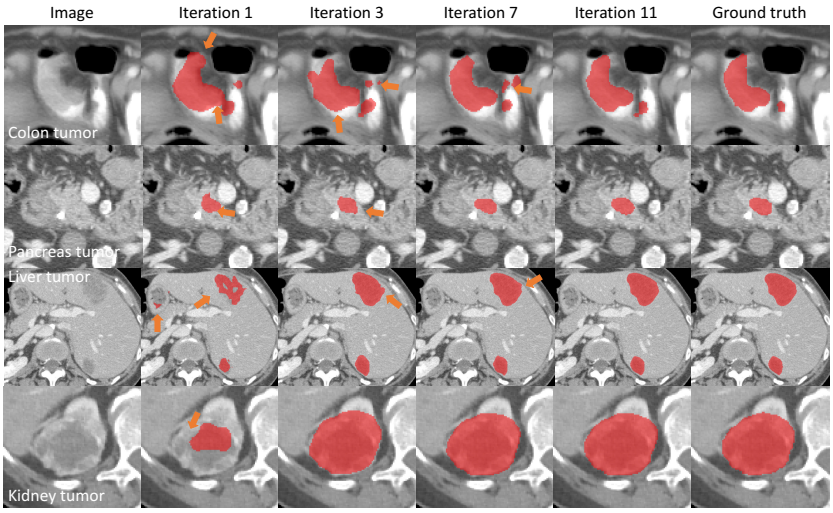

Fig. S.2: Qualitative results of PRISM-ultra across iterations. The orange arrows indicate the major defects which are corrected in the subsequent iteration.

Table S.1: Detailed settings for architecture and learning strategy in the proposed ablation studies.

	Encoder type			Learning strategy	
Variants	ViT	CNN	ViT + CNN	Confidence learning	Corrective learning
ViT encoder	✓				
CNN encoder		✓			
Hybrid			✓		
PRISM-plain			✓	✓	✓
PRISM-plain-b			✓	✓	✓
no ConL			✓		✓
PRISM-basic			✓	✓	✓
no CorL			✓	✓	
PRISM-ultra			✓	✓	✓

Table S.2: Detailed prompt setting for the proposed ablation studies. Tr. and T. represent training and test, respectively. The notation [1, 50] indicates the range from which the number of point prompts is randomly sampled during training. The term "varies" refers to the number of points used in tests, which include 1, 10, 50, and 100. "-erode" and "-dilate" denote box sizes that are 5 voxels smaller or larger in each dimension, respectively.

	Detailed prompt setting				
Variants	point	point num. (Tr. / T.)	box	box type in T.	scribble usage
ViT encoder	✓	1 / 1			
CNN encoder	✓	1 / 1			
PRISM-plain	✓	1 / 1			
PRISM-plain-b	✓	1 / 1	✓	tight	
no ConL	✓	1 / 1	✓	tight	
-plain-b-1	✓	1 / 1	✓	tight	
-plain-b-50	✓	50 / varies	✓	tight	
PRISM-basic	✓	[1, 50] / 1	✓	tight	
no CorL	✓	[1, 50] / 1	✓	tight	
-erode	✓	[1, 50] / varies	✓	undersized	
-dilate	✓	[1, 50] / varies	✓	oversized	
PRISM-ultra	✓	[1, 50] / varies	✓	tight	T.
PRISM-ultra+	✓	[1, 50] / varies	✓	tight	Tr. and T.