ASSESSING TEST-TIME VARIABILITY FOR INTERACTIVE 3D MEDICAL IMAGE SEGMENTATION WITH DIVERSE POINT PROMPTS

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

Vanderbilt University

ABSTRACT

Interactive segmentation model leverages prompts from users to produce robust segmentation. This advancement is facilitated by prompt engineering, where interactive prompts serve as strong priors during test-time. However, this is an inherently subjective and hard-to-reproduce process. The variability in user expertise and inherently ambiguous boundaries in medical images can lead to inconsistent prompt selections, potentially affecting segmentation accuracy. This issue has not yet been extensively explored for medical imaging. In this paper, we assess the test-time variability for interactive medical image segmentation with diverse point prompts. For a given target region, the point is classified into three subregions: boundary, margin, and center. Our goal is to identify a straightforward and efficient approach for optimal prompt selection during test-time based on three considerations: (1) benefits of additional prompts, (2) effects of prompt placement, and (3) strategies for optimal prompt selection. We conduct extensive experiments on the public Medical Segmentation Decathlon dataset for challenging colon tumor segmentation task. We suggest an optimal strategy for prompt selection during test-time, supported by comprehensive results. The code is publicly available at https://github.com/MedICL-VU/variability

***Index Terms*—** Medical image segmentation, test-time variability, interactive segmentation, point prompt selection, pretrained Segment Anything Model (SAM)

1. INTRODUCTION

To date, deep learning methods have demonstrated superior performance in various medical image segmentation tasks [1]. However, the generalizability and effectiveness of these fully automated methods may be limited by the amount of available labeled medical data [2]. Instead, interactive segmentation methods that integrate user knowledge and application requirements have been proposed [3].

To tackle the data and label availability issue in medical imaging, it is desirable to utilize knowledge from the natural images domain, wherein large public datasets are more readily accessible [4]. Recent studies [5, 6, 7, 8, 9, 10, 11] have leveraged insights from pretrained natural image foundation models, such as the Segment Anything Model (SAM) [12], which is trained on massive datasets, to facilitate robust medical image segmentation through parameter-efficient transfer learning techniques. These methods, which use diverse visual prompts, achieve robust performance due to the strong prior provided by the interactive prompt during test-time.

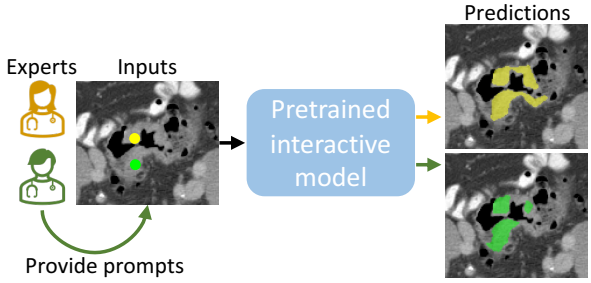

Fig. 1. A 3D interactive segmentation model, which takes an image and a point prompt as inputs. The segmentation result can vary with different prompts provided by experts.

Compared to other interaction formats, such as boxes or scribbles, point prompts are preferable in practice, as they require less effort, especially for 3D medical images. However, in the context of interactive segmentation models, determining precise key points during inference can be elusive, especially in medical images characterized by low quality, poor contrast, and ambiguous boundaries. Furthermore, subjectivity leads to variability in prompt choice as seen in Fig. 1, and this can be exacerbated by different user expertise levels, resulting in different segmentation outcomes at test-time [13, 14, 15]. A random selection strategy is widely used, but random points might not represent the key features of a medical image effectively. To the best of our knowledge, no existing work has explicitly addressed prompt selection in the medical field. Previous works have highlighted the importance of key points and have leveraged these during training to produce robust segmentations [16, 17]. In this paper, we aim to investigate the optimal strategy for point prompt selection for pretrained interactive segmentation models at test-time.

We first assess the test-time variability caused by point prompt selection of interactive 3D medical image segmentation model. Specifically, we use the ProMISe [8] model as

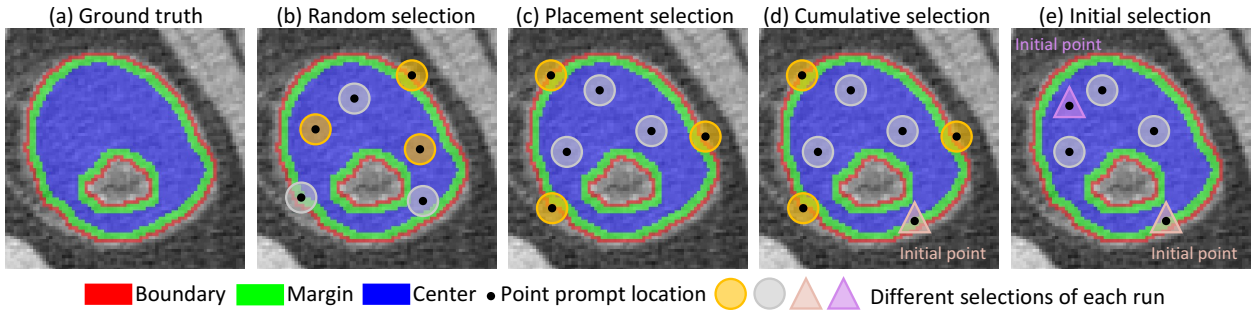

Fig. 2. Various point prompt selection strategies. Different prompt colors represent different runs. **(a)** Ground truth, highlighting the entire tumor and its three different sub-regions. **(b)** Random selection throughout the entire tumor. **(c)** Prompts confined to a specific sub-region. **(d)** Cumulative selection, where the initial point (triangle) is fixed while cumulative points (circles) vary between runs. **(e)** Initial selection, where initial points vary while cumulative points remain fixed.

the backbone, which leverages the pretrained weights from SAM [12]. It takes image and point prompts as inputs to produce a segmentation. We evaluate the segmentation performance variability caused by different number, region, and selection strategy for prompts during inference to quantify how such interactive models respond to diverse point prompts (Fig. 2). To provide a pragmatic solution, our investigation focuses on three aspects: (1) determining the necessary number of prompts, (2) identifying effective prompt placement locations, and (3) formulating a strategy for prompt selection.

We conduct our evaluation on colon tumor segmentation from the public Medical Segmentation Decathlon dataset [18], which is characterized by irregular shapes and ambiguous boundaries. Our findings suggest a straightforward strategy for test-time prompt selection without requiring additional effort, leading to significantly improved segmentation results over random prompt selection.

2. METHODS

Dataset. We used the Medical Segmentation Decathlon [18] for our experiments. We select the challenging colon tumor segmentation task where ambiguous edges and irregular shape are present. The dataset contains 126 3D CT volumes with resolution ranging from $0.54 \times 0.54 \times 1.25$ to $0.98 \times 0.98 \times 7.5mm^3$, resampled to $1mm$ isotropic resolution. The training, validation and test sets contain 88, 12 and 26 subjects, respectively. We perform intensity clipping based on foreground 0.5 and 99.5 percentiles, and Z-score normalization based on foreground voxels of training set.

Interactive segmentation model. We employ ProMISe [8], a SAM-based interactive segmentation model which takes an input image and point prompts as inputs. During training, an input patch of size $128 \times 128 \times 128$ was randomly selected such that its center voxel is equally likely to be foreground or background. In addition, random flip, rotation, zoom and intensity shift are used as data augmentation strategies. For point prompts, 10 positive points from the foreground and 20 negative points from the background are randomly selected for each input patch, respectively. We use negative points only if the number of positive points in an input patch is fewer than 10. During inference, ProMISe takes the input image and only a single random point prompt within the whole tumor area to produce the segmentation (Fig. 1). We consider this random prompt selection as the baseline method and evaluate various other point prompt selection strategies, aiming to identify a better strategy to optimize the segmentation performance without requiring extensive additional user effort.

Sub-region generation. As shown in Fig. 2(a), we divide the ground truth mask into three distinct sub-regions: boundary, margin, and center. We generate these regions using an average-pooling operation, denoted as P^k_{ave}, where k represents the kernel size. For a given a binary segmentation S, the binary boundary sub-region is derived: $B(S) = thres(S \odot |S - P^3_{ave}(S)|)$, where $|\cdot|$ denote the absolute value function, $\odot$ represents element-wise multiplication, and $thres$ is binary thresholding with a threshold of 0. Similarly, we obtain the margin sub-region as: $M(S) = thres(S \odot |S - P^7_{ave}(S)|) - B(S)$, and the center sub-region is simply $C(S) = S - M(S) - B(S)$. This sub-region generation can be seamlessly integrated into the inference without reducing computation speed, and may be used to facilitate pseudo-label learning.

Random selection. To ensure objectivity and fair comparison, random selection serves as the underlying method in our approach for every selection strategy, instead of actual user prompts. Additionally, we use randomly selecting single point prompts within the whole tumor area as the baseline method for comparison, as this represents the most common selection strategy. We specify the random seed to control variability and assess the impact of different settings within these selection strategies. It is important to note that the location of point prompts depends on the seed and the selected region. In other words, for a specific region and seed, the point locations remain the same if the number of point prompts is unchanged, and differences only arise with additional points. Moreover,

Table 1. Quantitative results of random selection, presented as $mean \pm std.dev.$ ○ and ● denote absent and present for the point prompt select region. **Bold** and underline denote the best performances for each **column (region-wise)** and row (prompt-wise), respectively. Blue indicates the baseline method.

Region			Dice score					Normalized surface Dice				
B	M	C	$1P$	$5P$	$10P$	$20P$	$100P$	$1P$	$5P$	$10P$	$20P$	$100P$
●	○	○	.622±.014	.650±.009	.652±.008	.650±.007	**.655±.006**	.768±.015	.798±.010	.803±.009	.801±.009	.806±.007
○	●	○	.630±.015	.652±.010	.654±.009	.652±.008	.654±.006	.776±.016	.802±.011	.805±.009	.805±.009	**.807±.008**
○	○	●	**.642±.012**	**.654±.006**	.654±.006	.652±.009	.653±.005	**.788±.015**	.802±.006	.804±.007	.803±.010	.805±.007
●	●	○	.632±.014	.653±.010	.652±.008	**.654±.007**	.652±.005	.777±.016	.803±.011	.802±.010	**.806±.007**	.804±.007
●	○	●	.634±.016	.652±.010	.653±.009	.653±.008	.652±.005	.779±.017	.801±.010	.803±.010	.804±.009	.805±.007
○	●	●	.634±.016	.654±.009	.654±.008	.653±.009	.654±.007	.779±.017	.803±.011	.804±.009	.805±.010	.807±.009
●	●	●	.637±.014	.653±.008	**.655±.008**	.653±.007	.652±.007	.783±.016	**.803±.008**	**.806±.009**	**.806±.007**	.804±.008

Abbreviations. B=boundary, M=margin, C=center, P=point prompt(s) per volume.

Table 2. Quantitative results of cumulative selection. **Bold** and underline denote the best performances for each **column (region-wise)** and row (prompt-wise), respectively. Orange indicates the suggested selection strategy.

Region				(Initial + cumulative) points							
$Init.$	$Cumu.$			Dice score				Normalized surface Dice			
W	B	M	C	$(1+4)P$	$(5+5)P$	$(10+10)P$	$(20+80)P$	$(1+4)P$	$(5+5)P$	$(10+10)P$	$(20+80)P$
●	●	○	○	.651±.011	.652±.008	.652±.007	**.655±.005**	.799±.009	.803±.009	.804±.008	**.807±.007**
●	○	●	○	.654±.011	.654±.008	.653±.008	.653±.005	.804±.009	**.806±.009**	.806±.009	.806±.007
●	○	○	●	**.657±.008**	**.654±.007**	**.655±.007**	.653±.007	**.805±.008**	.804±.008	**.808±.009**	.806±.008
●	●	●	●	.654±.010	**.654±.007**	.653±.007	.652±.007	.804±.010	.805±.008	.805±.008	.804±.008

Abbreviations. W=whole, B=boundary, M=margin, C=center, P=point prompt(s) per volume, $Init.$=Initial, $Cumu.$=cumulative.

we investigate the benefits of using more than one prompt. Specifically, we use 1, 5, 10, 20, and 100 prompts.

Placement selection. To identify the effectiveness of point placement, the point prompts are randomly selected within the given regional constraints. Fig. 2(c) depicts a situation with a same seed setting but different selection region.

Strategy selection. The initial selection is important in prompt selection [16]. To further examine its test-time variability, we assess two strategy selections, cumulative and initial selections, as shown in Figs. 2(d) and (e), respectively. For cumulative selection, we randomly set an initial point (triangle in Fig. 2(d)) in a specific region as the fixed point, and then compare the segmentation performance for different placements of the cumulative points (circles). All selected prompts, including initial and cumulative points, are used for each run. Similarly, in Fig. 2(e), we select different initial points while keeping the cumulative points fixed. These approaches are practical since selecting a single point prompt is preferred in practice for its minimal effort, compared to other prompt types such as boxes, scribbles or multiple point prompts. Furthermore, a preliminary segmentation or pseudo label may first be generated from the initial point(s) to serve as a reference for the selection of the cumulative points.

Implementation details. We trained the ProMISe [8] model for a maximum of 200 epochs with batch size of 1. The initial learning rate was 0.0004, decreasing by $2e^{-6}$ every epoch, and the AdamW optimizer was employed. During inference, we randomly selected a set of 50 random seeds and used it for each selection strategy. The Dice score and normalized surface Dice were used for evaluation, and we report mean and standard deviations over the 50 runs (seeds). We used PyTorch, MONAI and an NVIDIA A6000 for our experiments.

3. RESULTS

Quantitative results. Tab. 1 presents a detailed comparison of random selections, focusing mainly on two practical questions: whether more points are needed, and where to select them. The results clearly show that using more than a single prompt benefits segmentation in both metrics compared to a single point prompt. However, a distinct cutoff in improvement is observed at five point prompts per 3D volume, which suggests diminishing returns past that. Compared to the baseline method, choosing a single point prompt focusing on the center region often yields superior segmentation. As the number of points increases, the impact of choosing different regions becomes less pronounced.

Tab. 2 compares the different cumulative strategies with different cumulative point placement and different number of prompts. We found that randomly selecting a single point

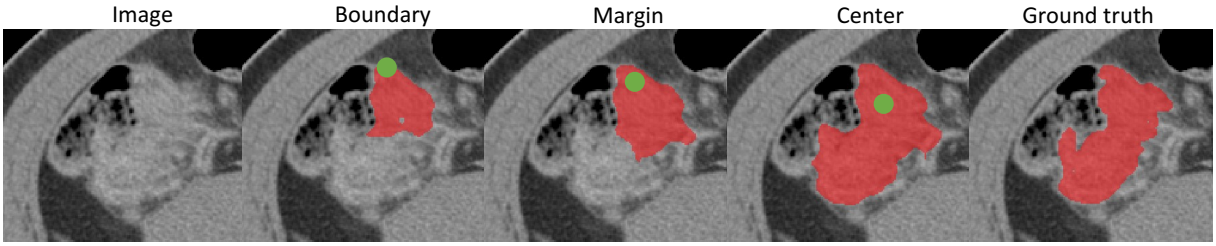

Fig. 4. Qualitative results for random selection with a single point prompt (green). Labels show the selection regions.

Table 3. Quantitative results of initial selection. The **bold** and underline denoted best performances for each **column (region-wise)** and row (prompt-wise), respectively. Orange indicates the suggested selection strategy.

		Region (Dice)			
Init.	*Cumu.*	*B*	*M*	*C*	*W*
$1(W)$	$4P$	.651±.011	.654±.011	**.657±.008**	.654±.010
$1(W)$	$9P$	.652±.008	.655±.008	**.655±.006**	.653±.010
$1(C)$	$4P$	.653±.008	.651±.007	**.654±.006**	.652±.008
$1(C)$	$9P$	.654±.007	.654±.007	**.654±.006**	.653±.009

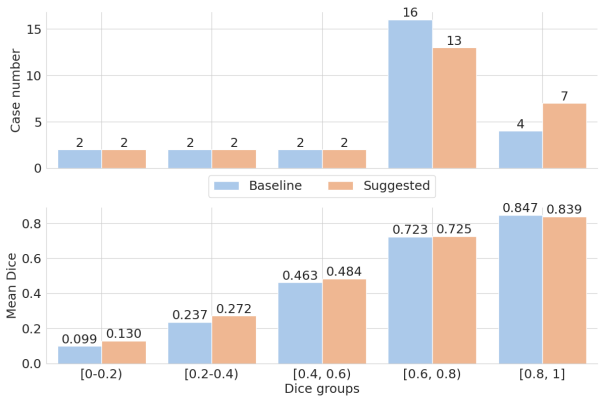

Fig. 3. Comparison of performance on Dice distribution.

prompt in the whole tumor area, with cumulative points picked from the center region achieves the best results. This improves the Dice score of baseline method by 2%, with statistical significance confirmed ($p < 0.001$) through a 2-tailed paired t-test. Therefore, we suggest this straightforward selection strategy as the optimal solution during test-time. We note that the cumulative points from the easily identifiable center area can either be derived from a pseudo label or require minimal extra effort by experts to improve segmentation performance. Thus, in practice, selecting a single initial point is enough for the recommended method, resulting in good performance for minimal prompt input effort.

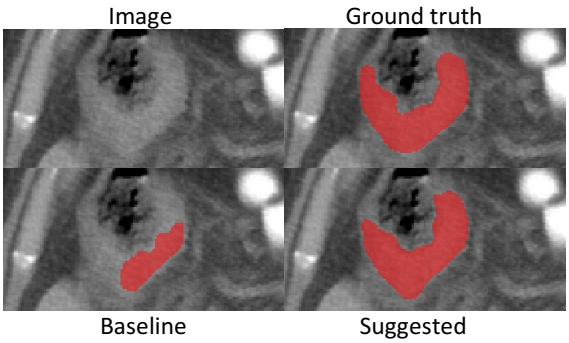

Fig. 5. Qualitative results of suggested and baseline methods.

Tab. 3 shows the impact of initial selection region. The results indicate that the whole and center areas are the best regions for initial and cumulative points, respectively. However, the differences among the methods are minor.

Fig. 3 presents the impact of suggested method on different subjects grouped by Dice. Compared to baseline method, the suggested method has higher Dice scores for most groups. In addition, the suggested method has more contribution to the subjects with high-quality segmentations produced by baseline method, as evidenced by Dice scores and case numbers for the two highest Dice groups.

Qualitative results. Fig. 4 shows qualitative results for the random selection with a single point prompt. Both boundary and margin selections produce under-segmented results near the ambiguous areas, while center selection captures the missing areas. In Fig. 5, the suggested method dramatically improves segmentation. The results match the ground truth well, with the exception of slightly over-segmented areas.

4. CONCLUSION

In this paper, we suggest a straightforward prompt strategy for interactive test-time segmentation, without requiring extensive additional effort. We evaluate on a public dataset for the challenging colon tumor segmentation task, with a significant improvement over the baseline method. Future work will validate the suggested method on more public datasets.

Acknowledgments. This work was supported, in part, by NIH U01-NS106845, and NSF grant 2220401.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by MSD. Ethical approval was not required as confirmed by the license attached with the open access data.

6. REFERENCES

[1] Han Liu, Dewei Hu, Hao Li, and Ipek Oguz, "Medical image segmentation using deep learning," in *Machine Learning for Brain Disorders*, pp. 391–434. Springer, 2023.

[2] Jiacheng Wang, Hao Li, Han Liu, Dewei Hu, Daiwei Lu, Keejin Yoon, Kelsey Barter, Francesca Bagnato, and Ipek Oguz, "Ssl2 self-supervised learning meets semi-supervised learning: multiple clerosis segmentation in 7t-mri from large-scale 3t-mri," in *Medical Imaging 2023: Image Processing*. SPIE, 2023, vol. 12464, pp. 126–136.

[3] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al., "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE transactions on medical imaging*, vol. 37, no. 7, pp. 1562–1573, 2018.

[4] Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith, "What makes transfer learning work for medical images: Feature reuse & other factors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9225–9234.

[5] Jun Ma and Bo Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023.

[6] 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.

[7] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou, "3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation," *arXiv preprint arXiv:2306.13465*, 2023.

[8] Hao Li, Han Liu, Dewei Hu, Jiacheng Wang, and Ipek Oguz, "Promise: Prompt-driven 3d medical image segmentation using pretrained image foundation models," *arXiv preprint arXiv:2310.19721*, 2023.

[9] Guoyao Deng, Ke Zou, Kai Ren, Meng Wang, Xuedong Yuan, Sancong Ying, and Huazhu Fu, "Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image," *arXiv preprint arXiv:2307.04973*, 2023.

[10] Yizhe Zhang, Tao Zhou, Peixian Liang, and Danny Z Chen, "Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model," *arXiv preprint arXiv:2304.11332*, 2023.

[11] Yichi Zhang and Rushi Jiao, "Towards segment anything model (sam) for medical image segmentation: A survey," 2023.

[12] 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.

[13] Arne Schmidt, Pablo Morales-Álvarez, and Rafael Molina, "Probabilistic modeling of inter-and intra-observer variability in medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21097–21106.

[14] Xing Yao, Han Liu, Dewei Hu, Daiwei Lu, Ange Lou, Hao Li, Ruining Deng, Gabriel Arenas, Baris Oguz, Nadav Schwartz, et al., "False negative/positive control for sam on noisy medical images," *arXiv preprint arXiv:2308.10382*, 2023.

[15] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li, "Sam on medical images: A comprehensive study on three prompt modes," *arXiv preprint arXiv:2305.00035*, 2023.

[16] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu, "Interactive image segmentation with first click attention," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13339–13348.

[17] Hong Joo Lee, Jung Uk Kim, Sangmin Lee, Hak Gu Kim, and Yong Man Ro, "Structure boundary preserving segmentation for medical image with ambiguous boundary," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4817–4826.

[18] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al., "The medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, pp. 4128, 2022.