ON THE ROBUSTNESS OF FOUNDATIONAL 3D MEDICAL IMAGE SEGMENTATION MODELS AGAINST IMPRECISE VISUAL PROMPTS

Soumitri Chattopadhyay, Başar Demir, Marc Niethammer

University of California, San Diego

ABSTRACT

While 3D foundational models have shown promise for promptable segmentation of medical volumes, their robustness to imprecise prompts remains under-explored. In this work, we aim to address this gap by systematically studying the effect of various controlled perturbations of dense visual prompts, that closely mimic real-world imprecision. By conducting experiments with two recent foundational models on a multi-organ abdominal segmentation task, we reveal several facets of promptable medical segmentation, especially pertaining to reliance on visual shape and spatial cues, and the extent of resilience of models towards certain perturbations. Codes are available at: https://github.com/ucsdbiag/Prompt-Robustness-MedSegFMs

***Index Terms*—** Prompt Robustness, Promptable foundational models, 3D medical image segmentation

1. INTRODUCTION

Interactive, or *promptable*, segmentation models [1, 2, 3, 4, 5] segment user-specified regions of interest, enabling general foreground–background separation beyond class-specific training. These models rely on ***visual prompts*** that are localized spatial cues that indicate the target structure. For natural images, SAM [1] supports point clicks, bounding boxes, and masks. Medical imaging adaptations extend these ideas: MedSAM [6] and SAM-Med2D [7] use bounding boxes in 2D, while SAM-Med3D [3], SegVol [4], and nnInteractive [5] employ clicks or combinations of multiple prompts (clicks, boxes, scribbles etc.) for 3D volumes. Many of these models refine predictions iteratively, using the previous segmentation as a ***dense mask prompt*** to guide subsequent updates [3, 5].

Despite these advances in promptable 3D medical image segmentation, performance of these models is *highly dependent on the quality of input prompts*. In practice, obtaining pixel-perfect visual prompts is often unrealistic: due to time constraints of high-quality annotation, lack of expert annotators, or atlas-driven prompt generation from imperfect registration [8, 9]. Yet, robustness to imprecise prompts is largely unexplored, despite growing clinical adoption.

In this work, we conduct a systematic probing of prompt robustness for 3D medical segmentation foundation models. For this study, we focus on ***dense visual prompts*** rather than manual sparse interactions (clicks, boxes, etc.), as they are *more flexible*: they enable automated prompt generation from auxillary segmentation [10], registration [8, 9] or supervoxels [11, 12], represent iterative refinement paradigm where *predictions become prompts* [3, 5], as well as can be used to simulate sparse interactions. From gold standard masks, we simulate realistic imprecise dense visual prompts through ***morphological perturbations*** and ***spatial translations***. Leveraging perturbed prompts, we experiment with two state-of-the-art foundational models, namely nnInteractive [5] and SAM-Med3D [3], on a multi-organ abdominal segmentation task [13]. Our controlled oracle experiments reveal ***(i)*** how models respond to varying prompt perturbations; ***(ii)*** how organ geometries relate to imprecise visual prompting; ***(iii)*** traits about their reliance on shape or boundary-aware priors compared to dense voxel-level precision.

To our knowledge, *this is the first robustness analysis of 3D medical segmentation foundation models against imprecise visual prompts.* While our experiments are synthetic, they mimic real-world scenarios where pixel-perfect prompts are unrealistic, making robustness evaluation timely and relevant.

2. MATERIALS AND METHODS

Models: We choose two recently proposed state-of-the-art foundational models for 3D medical segmentation – **nnInteractive** [5] and **SAM-Med3D** [3]. While the latter expands the SAM [1] model architecture to 3D volumes, nnInteractive uses a CNN backbone with prompts encoded in different channels of the input (along with the image).

Designing imprecise visual prompts: First, we perform morphological corruptions: ***(i)*** **dilation** and ***(ii)*** **erosion**, on the gold standard dense masks to yield perturbed coarse masks to be used as dense prompts. We vary the radius of dilation and erosion as integers in range $[1, 8]$ to obtain varying degrees of prompt coarseness. Dilated and eroded regions of interest typically mimic practical scenarios where voxel-level precision is often compromised for faster processing, lack of domain expertise, or automatically generated prompts e.g. registration [8] from an existing annotated image. Formally,

$$\mathcal{M}_{\texttt{dilate}} = \mathcal{M}_{\texttt{orig}} \oplus \mathcal{B}_r, \quad \mathcal{M}_{\texttt{erode}} = \mathcal{M}_{\texttt{orig}} \ominus \mathcal{B}_r \tag{1}$$

where $\oplus$ and $\ominus$ denote morphological dilation and erosion operations; $\mathcal{B}_r$ expands/contracts by r voxels in all directions, $r \in \{1, \ldots, 8\}$.

Furthermore, to disentangle the roles of boundary localization and dense-level context, we also include ***(iii)*** **boundary-preserving erosion**, a synthetic corruption that maintains the boundary outline of the mask while forming a cavity within its interior. This contrasts with standard erosion described above (which loses boundary information), but in turn lets us analyze the effect of shape-aware visual cues for promptable models. Formally, for boundary thickness $d \in \{1, \ldots, 8\}$, we have

$$\mathcal{M}_{\texttt{cavity}} = \mathcal{M}_{\texttt{orig}} \setminus (\mathcal{M}_{\texttt{orig}} \ominus \mathcal{B}_d) \tag{2}$$

Additionally, we also probe with ***(iv)*** **laterally translated masks** i.e. dense visual prompts which are spatially shifted from the true target region, again mimicking practical clinical usage when prompts are automated from unaligned images [8]. We apply small magnitudes of translation on all of the axes (without any morphological changes), since there is a large variation in sizes of organs and uncontrolled shifts may completely lose the region of interest in context.

$$\begin{aligned} \mathcal{M}_{\texttt{shift}}[x, y, z] = \mathcal{M}_{orig}[x + \delta_x, y + \delta_y, z + \delta_z] \\ \text{where } \delta_{\{x,y,z\}} \sim \mathcal{U}\{-3, \ldots, 3\} \end{aligned} \tag{3}$$

Having obtained the perturbed dense visual prompts, we follow the respective model workflows to use them as inputs and obtain segmentation predictions using the frozen models.

3. EXPERIMENTAL SETUP

Implementation. All source codes were implemented using PyTorch [14]. We used the publicly available code repositories as well as pre-trained released checkpoints of the respective foundational models to develop our dense-promptable inferencing setup. For nnInteractive[1] [5], all pre- and post-processing is handled within the model itself, hence we simply feed in the volumes with the augmented masks as dense prompts, all in their original resolutions. For SAM-Med3D[2] [3], the input volumes were first resampled to $1.5mm \times 1.5mm \times 1.5mm$ space, ROI cropped at $128 \times 128 \times 128$ around the organs, and Z-score normalized, following their exact preprocessing steps [3]. We conducted our experiments on a 48GB Nvidia RTX A6000 GPU.

Testing Dataset. We perform our evaluations on the Beyond The Cranial Vault **(BTCV)** Challenge [13], a multi-organ segmentation dataset comprising 30 abdominal CT volumes. The organs encompassed by the dataset include: *liver, kidneys, spleen, gallbladder, esophagus, adrenal glands, stomach, aorta, pancreas, inferior vena cava*, and the *portal and splenic vein* (i.e. 13 structures). The rationale for choosing this dataset is two-fold: ***(i)*** the *wide range of sizes of organs* covered in this dataset allows for reliable and insightful probing with imprecise visual prompts; ***(ii)*** this dataset was *not part of the training* of either of these foundational model checkpoints[3].

[1] https://github.com/MIC-DKFZ/nnInteractive

[2] https://github.com/uni-medical/SAM-Med3D

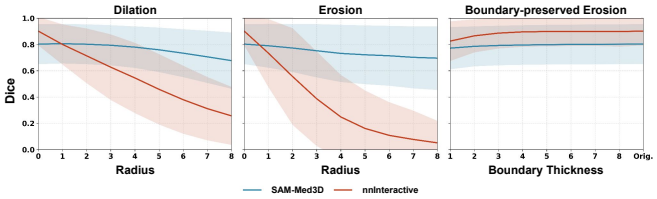

Fig. 1. Mean trends aggregated over all organs for different prompt perturbations. The respective color bands denote standard deviations across all samples. Models show greater resilience to dilated prompts as compared to eroded ones, as well as to prompts containing bounding shape information.

Metric. We report **Dice** score, a de facto standard metric for evaluating segmentation models [15, 16].

4. FINDINGS AND ANALYSIS

We highlight the key findings in terms of trends, and follow them up with discussions in the subsequent sections.

Models show greater robustness if the visual prompts contain boundary information.

In Figure 1, we show the aggregated global trends over all organs of BTCV, across varying strengths of perturbations – dilation, erosion and boundary-preserved erosion. We observe that while segmentation performance deteriorates with increasing coarseness across all of the perturbations, models show greater resilience when the structural boundary is preserved in the visual prompt compared to the other cases. Both dilated and eroded masks coarsely indicate the target region but fail to provide explicit spatial constraints on segmentation extent, leading to suboptimal segmentation. Conversely, boundary-preserving erosion maintains shape topology, effectively "telling the model" the bounds within which to segment – thereby injecting a stronger geometric prior and enabling greater robustness. We further provide qualitative examples to support this empirical observation in Figure 2, where clearly nnInteractive follows the dilated or eroded mask prompts and ends up segmenting beyond the true boundaries of structures, while it "fills up" the cavity of a boundary-informed visual prompt. Critically, boundary-preserving and fully eroded masks can have identical corruption levels (percentage of

[3] nnInteractive ckpt. trained on: https://www.codabench.org/competitions/5263/

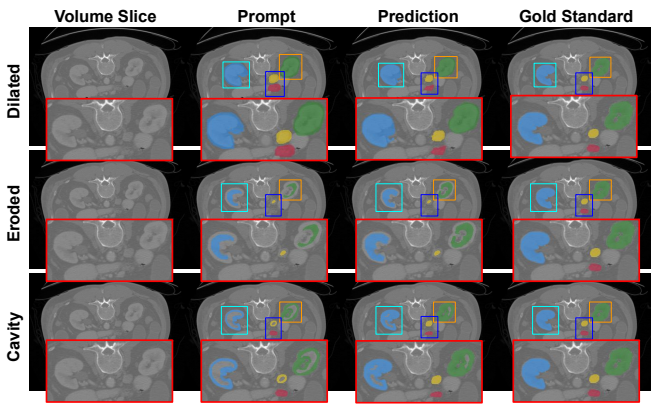

Fig. 2. Qualitative results of nnInteractive for different prompt perturbations. Inset shows zoomed-in regions containing the organs, their perturbed prompts, predictions and the gold standard. As can be seen, model over- and under-segments for dilated and eroded prompts, while it is more reliable when the prompt has bounding voxels ("Cavity").

pixels removed), yet models show greater robustness to the former, isolating the effect of preserved spatial geometry over dense voxel-level precision.

It should be noted that boundary-preserving erosion is a less-realistic simulation of imprecise prompting compared to other corruptions in clinical settings. However, we justify its inclusion given that we focus on synthetic simulations, and the empirical findings reveal a salient trait of promptable segmentation models – ***they benefit from shape descriptive visual prompts*** (since a boundary-preserving dense prompt provides a stronger shape prior compared to dilated/eroded ones).

Segmentation robustness to imprecise prompts varies greatly with geometry of the target structures.

We next examine organ-specific trends for each of these prompt perturbations. We choose representatives across the wide range of geometries of abdominal organs in BTCV [13]: *liver* (large); *spleen* and *kidney* (medium-sized), *gallbladder* (small), *pancreas* (irregular-shaped), and *aorta* (tubular shape). The trends are shown in Figure 4. We observe that eroded prompts are always worse than dilated prompts, although the geometry of the organs drives their resilience against these perturbations. For a bigger structure like liver, the segmentation performance deteriorates at a lower rate (Dice shows very little drop up to radius of dilation/erosion = 5), whereas for spleen, kidney and gallbladder (which are smaller structures), the drop is more aggressive even for smaller radii. Irregular and tubular structures (pancreas and aorta respectively) show greater rates of deterioration with increasing radius of dilation/erosion. As shown in the global trends, models are more robust to boundary-preserved erosion, where the shape and bounds of segmentation are present. The qualitative results of the prompts and the predicted segmentations in Figure 3 (as well as in Figure 2) further depict how the variation of organ geometry (and in turn, the prompts) degrades the performance of the segmentation in these organs. As for model trends, nnInteractive always yields superior segmentation with more precise prompts, but it degrades rapidly with increasing coarseness compared to SAM-Med3D (more discussion follows in a later section).

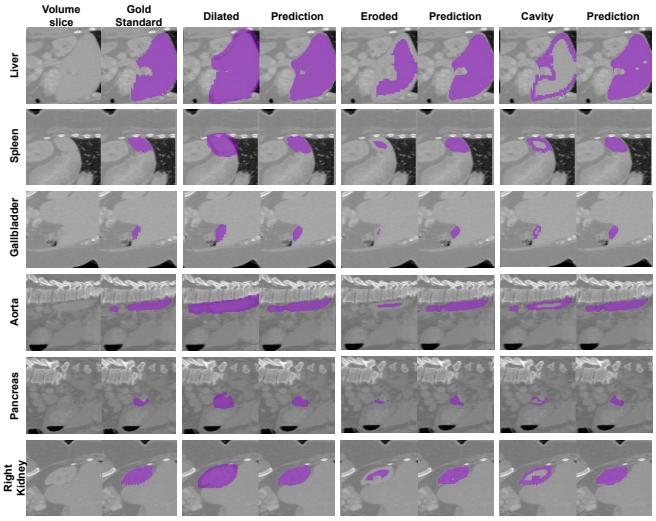

Fig. 3. Qualitative results of SAM-Med3D for different prompt perturbations. Each row depicts a different organ.

Bigger organs are more robust to small spatial shifts in dense prompts, while smaller organs suffer more.

We also test the promptable models with spatially misaligned (i.e. shifted) dense prompts for each organ. Figure 5 shows the empirical variations of SAM-Med3D *(top)* and nnInteractive *(bottom)*. We find SAM-Med3D to be robust to such spatial shifts for most organs. For nnInteractive, we observe that segmentation degradations are lower for regular shaped organs such as liver, kidneys and spleen, while more so in smaller organs (adrenal glands, gallbladder) and organs with irregular shape (portal and splenic vein, esophagus). A fundamental difference between spatial translation and morphological operations is that the former preserves the original geometry of the target structure, making the prompt geometry-aware (similar to boundary-preserving erosion), and so the models can better recover the original organ from the spatially shifted input compared to dilation/erosion. We also show qualitative samples in Figure 6 for various organs. Combining with earlier observations from boundary-preserved eroded prompts, we hypothesize that *shape-aware visual prompts are stronger than mere spatial location-based visual cues* to drive promptable 3D segmentation.

nnInteractive vs. SAM-Med3D: Broader Discussion

While we show and analyze trends between the two foun-

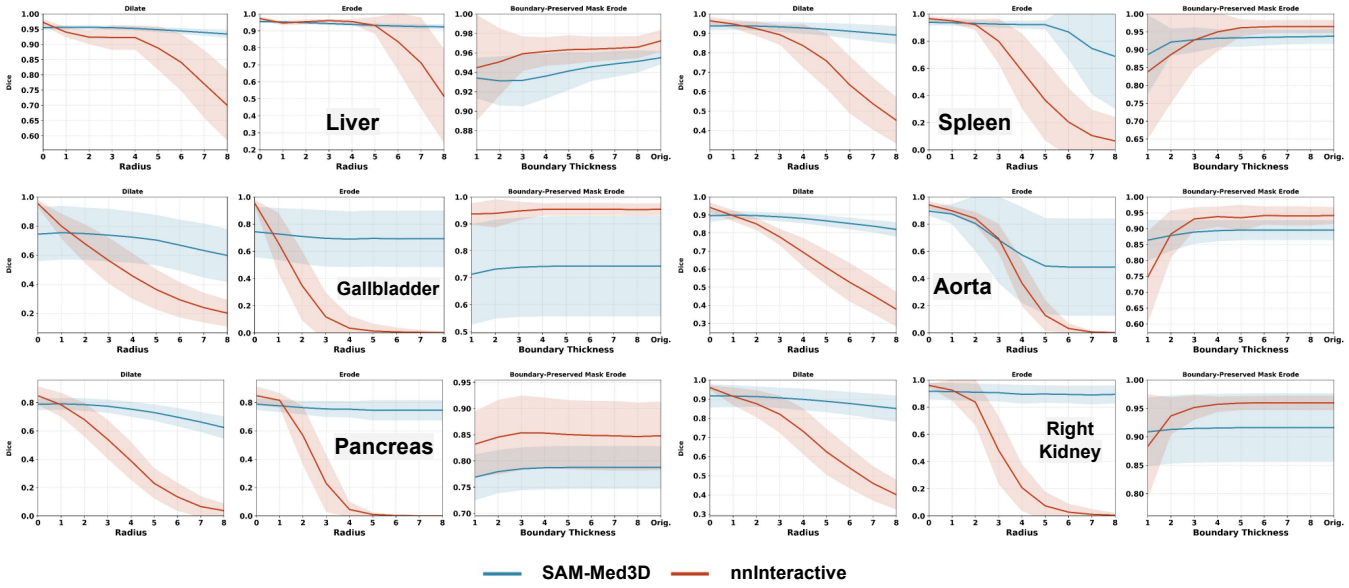

Fig. 4. Segmentation trends across varying strengths of prompt perturbations (radius for dilation/erosion, boundary thickness for boundary-preserved erosion) for different organs. The respective color bands depict the standard deviation across all samples.

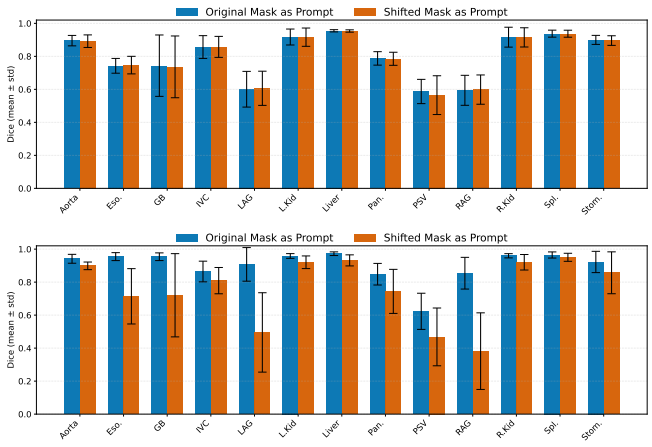

Fig. 5. Organ-wise performance for spatially shifted prompts, for SAM-Med3D *(top)* and nnInteractive *(bottom)*.

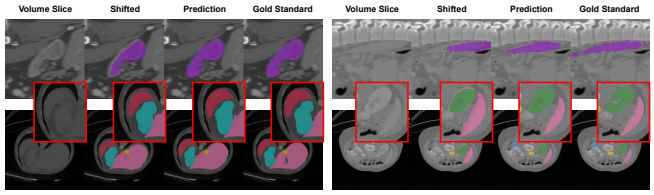

Fig. 6. Qualitative visuals depicting spatially shifted prompts (zoomed-in) and their respective predictions for abdominal organs, using SAM-Med3D *(top)* and nnInteractive *(bottom)*.

dational segmentation models nnInteractive [5] and SAM-Med3D [3] across perturbed prompts, it is also essential to include a broader discussion about the models themselves. Firstly, nnInteractive injects prompts at original resolution (i.e. in separate channels along with the input image), while SAM-Med3D follows the SAM [1] paradigm and injects prompts at downsampled features. Secondly, SAM-Med3D applies a 128^3 hard cropping and resizing around the target organ, which is fed into the model. While this allows for faster processing, it is also not realistic to expect a user to crop the ROI every time. Secondly, the ROI cropping allows the target region to be a major part the field of view at all times, which may contribute to the relatively higher stability of SAM-Med3D towards prompt perturbations compared to nnInteractive, which contrarily processes all data at native resolution. However, nnInteractive applies a computationally expensive auto-zoom mechanism that adaptively expands the ROI to include the entirety of the organ (please refer to [5] for in-depth understanding), which slows down the overall processing speed in return for higher precision. Thirdly, from Figure 5, *without any perturbation to the prompts, nnInteractive significantly outperforms SAM-Med3D on all organs.*

5. CONCLUSION

We explored the sensitivity of 3D promptable medical segmentation models to varying imprecision in visual prompts. Using controlled synthetic experiments mimicking real-world unavailability of precise visual prompts, we showed existing models generally struggle to overcome perturbed prompts, often over- or under-segmenting target structures, especially for smaller or irregular-shaped organs. Through boundary-preserved erosion simulation, we revealed the usefulness of prominent contextual shape information rather than dense textural cues, which may lead to new directions in training prompt-based segmentation models. In future, we will leverage these observations and develop robust training paradigms for reliable visually promptable segmentation models.

6. ACKNOWLEDGEMENTS

This research was, in part, funded by the National Institutes of Health (NIH) under other transactions 1OT2OD038045-01 and NIAMS 1R01AR082684. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the NIH.

7. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access from the datasets aptly cited in our paper. Ethical approval was not required as confirmed by the respective licenses attached with the open access data.

8. REFERENCES

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al., "Segment anything," in *ICCV*, 2023.

[2] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu, "A comprehensive survey on segment anything model for vision and beyond," *arXiv preprint arXiv:2305.08196*, 2023.

[3] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, et al., "SAM-Med3D: A vision foundation model for general-purpose segmentation on volumetric medical images," *IEEE TNNLS*, 2025.

[4] Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao, "Segvol: Universal and interactive volumetric medical image segmentation," in *NeurIPS*, 2024.

[5] Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, et al., "nnInteractive: Redefining 3d promptable segmentation," *arXiv preprint arXiv:2503.08373*, 2025.

[6] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang, "Segment anything in medical images," *Nature Communications*, 2024.

[7] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, et al., "Sam-med2d," *arXiv preprint arXiv:2308.16184*, 2023.

[8] Lin Tian, Hastings Greer, Roland Kwitt, François-Xavier Vialard, et al., "unigradicon: A foundation model for medical image registration," in *MICCAI*, 2024.

[9] Basar Demir and Marc Niethammer, "Multimodal image registration guided by few segmentations from one modality," in *MIDL*, 2024.

[10] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, 2021.

[11] Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, et al., "Vista3d: Versatile imaging segmentation and annotation model for 3d computed tomography," in *CVPR*, 2025.

[12] Pedro F Felzenszwalb and Daniel P Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, 2004.

[13] Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, 2015.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, et al., "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.

[15] Lee R Dice, "Measures of the amount of ecologic association between species.," *Ecology*, 1945.

[16] Soumitri Chattopadhyay, Basar Demir, and Marc Niethammer, "Zero-shot domain generalization of foundational models for 3d medical image segmentation: An experimental study," *arXiv preprint arXiv:2503.22862*, 2025.