Title

Necessity and impact of specialization of large foundation model for medical segmentation tasks.

Authors

Nguyen, Eric
Liu, Hengjie
Ruan, Dan

Publication Date

2024-10-21

DOI

10.1002/mp.17470

Peer reviewed

Received: 8 May 2024 | Revised: 17 August 2024 | Accepted: 30 September 2024

DOI: 10.1002/mp.17470

RESEARCH ARTICLE

MEDICAL PHYSICS

Necessity and impact of specialization of large foundation model for medical segmentation tasks

Eric Nguyen[1] | **Hengjie Liu**[1] | **Dan Ruan**[1,2]

[1]Department of Radiation Oncology, University of California Los Angeles, Los Angeles, California, USA

[2]Department of Bioengineering, University of California Los Angeles, Los Angeles, California, USA

Correspondence
Dan Ruan, Department of Radiation Oncology, University of California Los Angeles, Los Angeles, California 90095, USA.
Email: druan@mednet.ucla.edu

Funding information
UCLA Council of Research

Abstract

Background: Large foundation models, such as the Segment Anything Model (SAM), have shown remarkable performance in image segmentation tasks. However, the optimal approach to achieve true utility of these models for domain-specific applications, such as medical image segmentation, remains an open question. Recent studies have released a medical version of the foundation model MedSAM by training on vast medical data, who promised SOTA medical segmentation. Independent community inspection and dissection is needed.

Purpose: Foundation models are developed for general purposes. On the other hand, stable delivery of reliable performance is key to clinical utility. This study aims at elucidating the potential advantage and limitations of landing the foundation models in clinical use by assessing the performance of off-the-shelf medical foundation model MedSAM for the segmentation of anatomical structures in pelvic MR images. We also explore the simple remedies by evaluating the dependency on prompting scheme. Finally, we demonstrate the need and performance gain of further specialized fine-tuning.

Methods: MedSAM and its lightweight version LiteMedSAM were evaluated out-of-the-box on a public MR dataset consisting of 589 pelvic images split 80:20 for training and testing. An nnU-Net model was trained from scratch to serve as a benchmark and to provide bounding box prompts for MedSAM. MedSAM was evaluated using different quality bounding boxes, those derived from ground truth labels, those derived from nnU-Net, and those derived from the former two but with 5-pixel isometric expansion. Lastly, LiteMedSAM was refined on the training set and reevaluated on this task.

Results: Out-of-the-box MedSAM and LiteMedSAM both performed poorly across the structure set, especially for disjoint or non-convex structures. Varying prompt with different bounding box inputs had minimal effect. For example, the mean Dice score and mean Hausdorff distances (in mm) for obturator internus using MedSAM and LiteMedSAM were $\{0.251 \pm 0.110, 0.101 \pm 0.079\}$ and $\{34.142 \pm 5.196, 33.688 \pm 5.306\}$, respectively. Fine-tuning of LiteMedSAM led to significant performance gain, improving Dice score and Hausdorff distance for the obturator internus to 0.864 ± 0.123 and 5.022 ± 10.684, on par with nnU-Net with no significant difference in evaluation of most structures. All segmentation structures benefited significantly from specialized refinement, at varying improvement margin.

Conclusion: While our study alludes to the potential of deep learning models like MedSAM and LiteMedSAM for medical segmentation, it highlights the need for specialized refinement and adjudication. Off-the-shelf use of such large foundation models is highly likely to be suboptimal, and specialized fine-tuning is often necessary to achieve clinical desired accuracy and stability.

KEYWORDS
foundation model, medical image segmentation

1 | INTRODUCTION

Foundation models are general models trained on extremely large and diverse datasets that make up a wide range of categories which enable the model to be multi-purpose.[1] Segment Anything (SAM) is a generalist foundation model developed by Meta for image segmentation. It is trained on over 11 million natural images with 1 billion masks and has shown remarkable zero-shot segmentation performance on a diverse range of image segmentation tasks.[2,3] Foundation models can potentially generalize across different imaging modalities, anatomical structures, and pathologies. This reduces the need for developing and maintaining multiple task-specific models. SAM utilizes a variety of prompts including point-prompting, bounding box, or auto-prompting to facilitate segmentation decoding.

However, studies have shown that SAM may be challenged by low image contrast and amorphous target structures in medical image segmentation, resulting in poor and unstable performance compared to natural image applications.[4–7]

Multiple methods have been devised to bridge this gap such as the incorporation of Low Rank Adaptation (LoRA) into SAM's image encoder.[4,8,9] Another method, SAM-Adapter, was developed primarily for improved detection of shadows and camouflaged objects in natural images, but has also been shown to improve performance of polyp segmentation for medical images.[10,11] Another study proposed a 3D adaptation of SAM, originally limited to 2D images, to better facilitate medical images.[5]

Ma et al. chose to preserve the architecture of the original SAM and refine the same foundation model with an unprecedented dataset with over a million annotated medical images, giving rise to the medical "version" of SAM, known as MedSAM.[12] MedSAM's training spans across diverse modalities including CT, MRI, endoscopy, ultrasound, X-ray, pathology, fundus photography, dermoscopy, and OCT. The authors report that MedSAM shows significant improvement over SAM for medical applications and can perform on par with modality-specific state-of-the-art models.

However, it is also important to note that rather than aiming for generality, solid clinical utilization prioritizes stability and consistent high accuracy in addressing a specific problem in a specific context. While it is tempting to use large foundation models off-the-shelf, whether that being the original SAM or the modified MedSAM, the question remains as to whether it is truly ready to deliver clinical value.

In this study, we perform an independent investigation and assessment of the clinical feasibility of the MedSAM foundation model on a new segmentation task, MR-based pelvic segmentation, using various variants in the MedSAM family. Specifically, we evaluate the performance of off-the-shelf MedSAM, its dependency and sensitivity on the prompt input, and impact of fine tuning. Comparison was performed against a special-purpose in-house trained nnU-Net benchmark.

2 | MATERIAL AND METHODS

2.1 | Dataset description

We used a dataset curated by Li et al. consisting of 589 T2w MRI images acquired from seven studies (INDEX, the SmartTarget Biopsy Trial, PICTURE, TCIA Prostate3T, Promise12, TCIA ProstateDx and the Prostate MR Image Database).[13] All the studies are represented equally in the total curated dataset. The collection contains images from multiple institutions that were done using scanners with either 1.5T or 3.0T field strengths and from two different manufacturers. The images had an in-plane resolution ranging from 0.3 mm to 1.0 mm and a slice thickness ranging from 1.8 mm to 5.4 mm and had varying field-of-views. All images distributed by the curated dataset had dimensions of Nx180 × 180 pixels. Manually annotated segmentations were also provided for each of the images. Eight anatomical structures were labelled, including bladder, femur bone, obturator internus, transition zone, central gland, rectum, seminal vesicle, and neurovascular bundle. Prostate and bone segmentation were included as part of MedSAM's original training data and task, but some of the other structures were not trained explicitly in MedSAM.

The current SAM and MedSAM can only cope with 2D segmentation, so slices in each 3D volume are processed and segmented independently and then recompiled back into the volume space for performance evaluation.

2.2 | Background on SAM and MedSAM

MedSAM consists of a vision transformer (ViT)-based encoder for feature extraction, a prompt encoder that accepts user-provided bounding box inputs, and a lightweight mask decoder. MedSAM uses image inputs of size 1024 × 1024 while LiteMedSAM uses image inputs of size 256 × 256. Our images were resampled to meet the respective dimension criteria.

MedSAM was pre-trained using a loss function defined as the unweighted sum of Dice loss and binary cross entropy (BCE) loss. Meanwhile, the loss function used for the pre-training of LiteMedSAM also includes intersection over union (IoU) loss into this summation. The loss function used by LiteMedSAM is shown below. Let N be the total number of voxels in an image and let g_i and s_i be the ith voxel of the ground truth and predicted segmentations, respectively.

$$L_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N} g_i \log s_i + (1-g_i)\log(1-s_i)$$

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=0}^{N} g_i s_i}{\sum_{i=0}^{N} g_i^2 + \sum_{i=0}^{N} s_i^2}$$

$$L_{\text{IoU}} = 1 - \frac{\sum_{i=0}^{N} g_i s_i}{\sum_{i=0}^{N} g_i^2 + \sum_{i=0}^{N} s_i^2 - \sum_{i=0}^{N} g_i s_i}$$

$$L = L_{\text{Dice}} + L_{\text{BCE}} + L_{\text{IoU}}$$

We examined the performance of the pre-trained MedSAM and a pre-trained LiteMedSAM on the testing data and task described in Section 2.4.

2.3 | Prompting variations

MedSAM models take bounding box input as prompts. We have conducted investigation on four different options of bounding box prompts: oracle bounding boxes derived from the ground truth mask without and with 5-pixel isometric expansion, and bounding boxes derived from nnU-Net segmentation mask without and with 5-pixel isometric expansion. The prompts derived from ground truth labels are used to reflect the best-case prompting while those derived from nnU-Net labels simulate average-case prompting.

2.4 | Refinement of LiteMedSAM with specialized data and task

For fine-tuning of LiteMedSAM, the original 589 3D MR images were randomly split at an 80:20 ratio for training and hold-off testing in this study. The training group was again randomly split at an 80:20 ratio for training and validation. In total, 18,350 2D image-mask pairs were used to train the LiteMedSAM model. Data augmentation was utilized during training which consisted of random left-right or up-down flips. During training, bounding boxes were also generated from the ground-truth masks and used as an additional model input. The bounding box inputs had 5-pixel random shift. The network was optimized using Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) with a learning rate of 5e-5 and a weight decay of 0.01. The same loss function used for training of LiteMedSAM was again used here for fine-tuning.

For consistency with the preprocessing protocol described in the MedSAM paper, the following two exclusion criteria were applied on the training dataset. To improve dataset quality, tiny objects defined by a 100-pixel threshold were removed from slices as the challenge would be detection rather than segmentation. In addition, the intensity levels of all images were clipped between the 0.5th and 99.5th percentiles, and min-max normalization was applied.

Training was performed on a single GPU (NVIDIA GeForce RTX3080, 10G memory) using resampled image inputs of size 256 × 256 to accommodate the memory limit. The model typically converges in about 300 epochs with a batch size of 4. The checkpoint with the best validation loss was chosen for the final model.

2.5 | Benchmark nnU-Net model

We trained a special-purpose in-house nnU-Net model to serve two purposes: a benchmark for overall performance comparison, and a means to generate bounding box prompts for MedSAM. We opted to use nnU-Net because it has been shown to perform well on a large variety of segmentation tasks.[14–16]

Our nnU-Net was trained from scratch and using the same dataset and dataset splitting as with LiteMedSAM fine-tuning. nnU-Net offers both 2D and 3D U-Net configurations for training on new datasets. We opted to use the "3D Full res" configuration which uses the native image resolution (Nx180 × 180). The nnU-Net model adaptively determines multiple hyperparameters by analyzing the specific characteristics of the dataset. The default network architecture is a U-Net consisting of 6 convolution blocks in the encoder and decoder. Each block consists of two convolution layers followed by batch normalization and a leaky ReLU activation function. Convolution steps were done with a 3 × 3 × 3 kernel size and stride length of 2. The patch size is 48 × 192 × 192 (zero-padding was automatically applied as per nnU-Net standard operation) and a batch size of 2 was used. Training was performed using an SGD optimizer with a learning rate of 0.01, momentum of 0.99, and weight decay of 3e-5. The default loss function for nnU-Net, a combination of Dice loss and cross-entropy loss, was used. The network was trained over 1000 epochs and the checkpoint with the best validation loss was chosen.

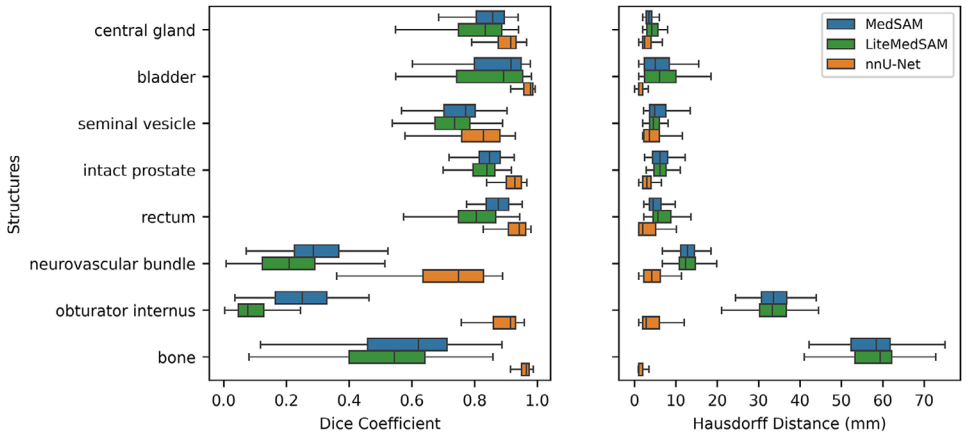

FIGURE 1 Comparison of Dice scores and Hausdorff distances among MedSAM (blue), LiteMedSAM (green), and nnU-Net (orange). While MedSAM and LiteMedSAM showcase comparable performance, nnU-Net is significantly superior.

2.6 | Evaluation metrics

Dice scores and 95% Hausdorff distances were calculated for each anatomical structure segmented for all patients using each of the models. Comparisons between each of the models were performed using a two-sided paired t-test on statistics generated from the testing of 117 volumetric images. A significance level of 0.05 was used.

3 | RESULTS

3.1 | Suboptimal performance for off-the-shelf SAM models

Regardless of the intrinsic segmentation challenge level based on the scale and morphology of the structure or intensity context, MedSAM or LiteMedSAM is notably inferior compared to the specialized nnU-Net performance, as shown in Figure 1. The benchmark nnU-Net showed statistically significant superiority in Dice score for all structures ($p < 0.0001$) and in Hausdorff distance for all structures ($p < 0.05$) except seminal vesicle in which no significance difference was noted. Specifically, for structures with well-defined and reasonably convex shapes, as in bladder, rectum, and intact prostate, MedSAM resulted in dice scores of 0.808, 0.846, and 0.833, respectively while LiteMedSAM achieved comparable dice scores of 0.783, 0.787, and 0.818, respectively. That's 10%–20% lower than mean dice scores of 0.958, 0.924, and 0.918 from the benchmark nnU-Net. In addition, both MedSAM and LiteMedSAM experienced great difficulty segmenting disjoint, bilateral structures including femur, neurovascular bundle, and obturator internus. For these structures, both models yielded mean Dice scores lower than 0.6. The specific values are reported in columns 1, 5, and 7 in Table 1. Illustrative examples are shown in Figure 2.

Our in-house nnU-Net performs as expected compared to its usage in other studies which suggests that it is a reliable reference benchmark. Bhandary et al. reported that nnU-Net achieved mean Dice scores of 0.850, 0.876, and 0.910 of the prostate when trained and evaluated on three different MR prostate datasets.[17] This is fairly consistent with our in-house nnU-Net which scored 0.918 for prostate segmentation but which had been trained on a larger compilation of datasets.

3.2 | Impact of bounding box prompt

MedSAM models currently only offer stable support for prompts in the form of bounding boxes. We evaluated the stability of MedSAM segmentations when provided with different bounding box prompts. To simulate best-case prompting, we provided MedSAM with oracle bounding boxes derived from the ground truth labels. Meanwhile, bounding boxes generated from nnU-Net was used to approximate average-case prompting. As shown in Figure 3, when provided bounding boxes generated by nnU-Net labels, MedSAM inferences saw a statistically significant decrease in mean Dice score compared to MedSAM inferences using ground-truth bounding boxes for prostate central gland (0.808 vs. 0.800), intact prostate (0.836 vs. 0.801), seminal vesicle (0.718 vs. 0.637), and neurovascular bundle (0.294 vs. 0.249). No significant difference in Hausdorff distances were observed between the two tests, except for seminal vesicle (6.34 mm vs. 8.23 m m).

As shown in Table 1 columns 3 and 4, isometric expansion of both the ground truth derived bounding boxes and the nnU-Net derived bounding boxes by 5 pixels showed a very minor increase in mean Dice score for all structures but were only found to be statistically significant in the case of central gland, neurovascular bundle, and obturator internus. No significant differences in Hausdorff distances were observed.

TABLE 1 Mean and standard deviation Dice Scores and 95% Hausdorff Distances for different models. Best value for each structure is bolded. Abbreviations: OI (obturator internus), NVB (neurovascular bundle), SV (seminal vesicle).

Model (bbox option)	MedSAM (oracle)	MedSAM (Unet)	MedSAM (oracle expanded)	MedSAM (Unet expanded)	LiteMedSAM (oracle expanded)	LiteMedSAM refined (oracle)	nnU-Net (none)
Dice Scores							
Bladder	0.81 ± 0.24	0.80 ± 0.24	0.81 ± 0.24	0.80 ± 0.24	0.78 ± 0.25	0.92 ± 0.17	**0.96 ± 0.06**
Bone	0.58 ± 0.19	0.57 ± 0.19	0.59 ± 0.18	0.58 ± 0.19	0.51 ± 0.19	0.92 ± 0.14	**0.95 ± 0.05**
Central gland	0.84 ± 0.10	0.80 ± 0.09	0.87 ± 0.09	0.83 ± 0.09	0.80 ± 0.12	**0.91 ± 0.08**	0.90 ± 0.04
Intact prostate	0.83 ± 0.09	0.81 ± 0.08	0.84 ± 0.09	0.82 ± 0.08	0.82 ± 0.09	0.90 ± 0.08	**0.92 ± 0.04**
NVB	0.29 ± 0.10	0.25 ± 0.10	0.33 ± 0.10	0.28 ± 0.10	0.22 ± 0.12	**0.72 ± 0.13**	0.71 ± 0.15
OI	0.25 ± 0.11	0.24 ± 0.11	0.29 ± 0.11	0.28 ± 0.11	0.10 ± 0.08	0.86 ± 0.12	**0.89 ± 0.06**
Rectum	0.85 ± 0.12	0.83 ± 0.11	0.86 ± 0.11	0.84 ± 0.11	0.79 ± 0.12	0.91 ± 0.11	**0.92 ± 0.05**
SV	0.72 ± 0.16	0.64 ± 0.17	0.74 ± 0.16	0.65 ± 0.18	0.69 ± 0.16	**0.81 ± 0.18**	0.79 ± 0.12
Hausdorff Distances							
Bladder	6.63 ± 6.05	6.68 ± 5.50	6.55 ± 6.17	6.55 ± 5.54	7.53 ± 6.44	2.76 ± 4.70	**1.97 ± 2.11**
Bone	56.15 ± 13.87	57.34 ± 13.13	55.80 ± 14.09	56.98 ± 13.34	57.37 ± 12.52	3.97 ± 14.11	**2.77 ± 11.53**
Central gland	4.12 ± 3.20	4.53 ± 2.51	3.67 ± 3.27	4.12 ± 2.52	4.74 ± 2.94	**2.64 ± 2.78**	3.17 ± 1.67
Intact prostate	6.82 ± 4.12	6.51 ± 3.02	6.83 ± 4.11	6.34 ± 2.35	6.75 ± 3.88	5.57 ± 4.58	**3.49 ± 2.85**
NVB	13.55 ± 5.35	13.43 ± 5.73	13.56 ± 5.34	13.39 ± 5.60	13.52 ± 5.47	5.72 ± 6.44	**5.49 ± 6.30**
OI	34.14 ± 5.20	33.66 ± 4.30	34.01 ± 5.33	33.50 ± 4.41	33.69 ± 5.31	5.02 ± 10.68	**4.98 ± 5.13**
Rectum	8.39 ± 12.47	8.25 ± 8.14	8.26 ± 12.62	8.21 ± 8.20	9.48 ± 11.99	6.03 ± 12.56	**5.10 ± 7.11**
SV	6.40 ± 4.31	8.23 ± 6.50	6.30 ± 4.39	8.17 ± 6.65	5.52 ± 3.42	**3.39 ± 3.19**	4.95 ± 4.96

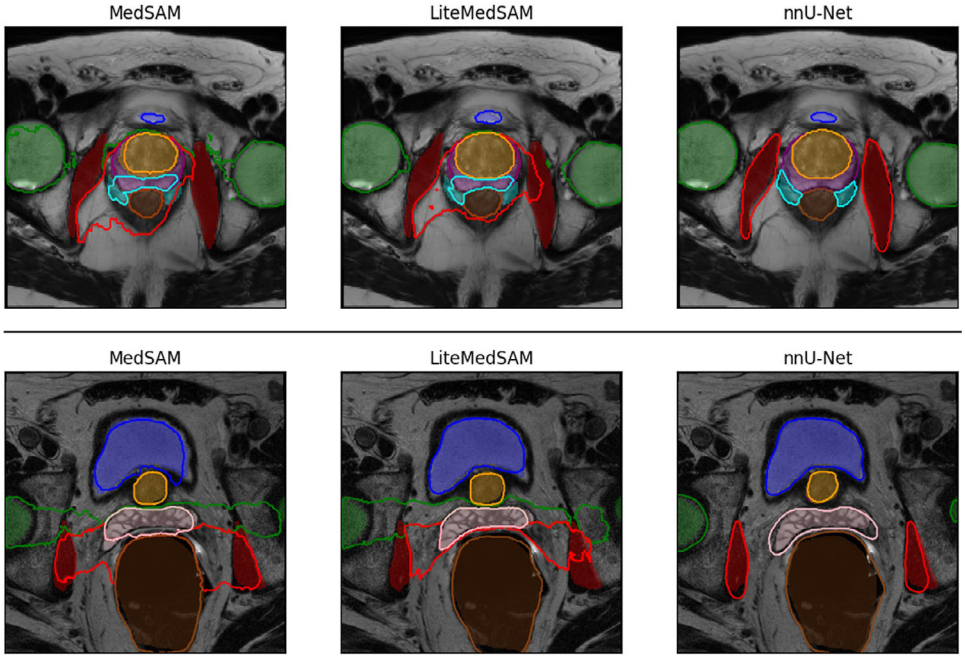

FIGURE 2 Examples of off-the-shelf MedSAM and LiteMedSAM segmentations compared to nnU-Net for two different patients. Shading indicates ground-truth. Structures shown: femur (green), bladder (blue), central gland (orange), intact prostate (purple), obturator internus (red), seminal vesicle (pink), neurovascular bundle (cyan), rectum (brown).

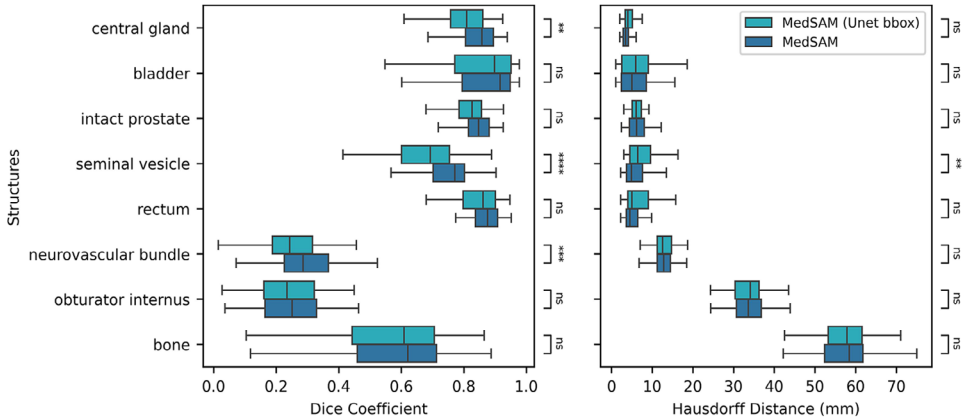

FIGURE 3 Comparison of MedSAM sensitivity to bounding box prompts derived by nnU-Net or ground truth labels. [*, **, ***, ****, ns] corresponds to statistical significance of $\alpha = [0.05, 0.01, 0.001, 0.0001, \text{No Significance}]$.

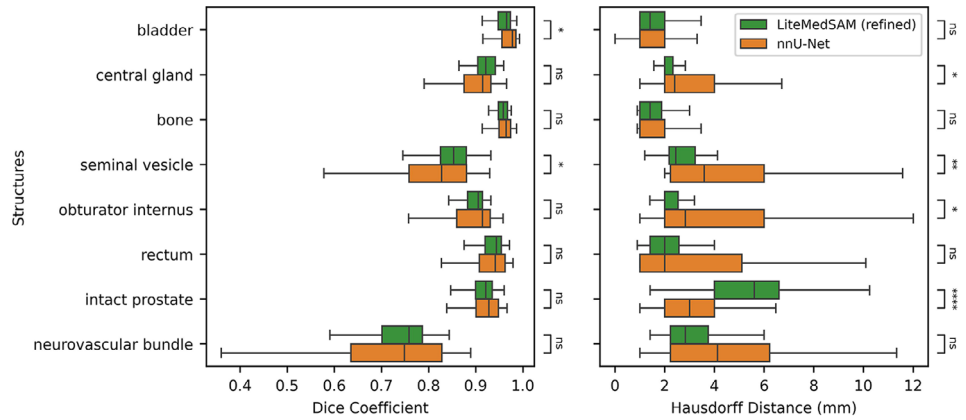

FIGURE 4 Comparison of Dice scores and Hausdorff distances between refined LiteMedSAM and nnU-Net. [*, **, ***, ****, ns] corresponds to statistical significance of $\alpha = [0.05, 0.01, 0.001, 0.0001, \text{No Significance}]$.

3.3 | Impact of specialized refinement

After being fine-tuned using a subset of our total Pelvic MR dataset, the LiteMedSAM model can perform on par with the benchmark nnU-Net, as shown in Figure 4. Whereas out-of-the-box LiteMedSAM struggled with disjoint object segmentation, the fine-tuned model has successfully learned to segment these structures more effectively. While most structures showed comparable performance between the refined LiteMedSAM model and the benchmark nnU-Net in either mean Dice score or mean Hausdorff distance, LiteMedSAM showed advantage in segmenting the seminal vesicle better. Figure 5 shows the improvement of LiteMedSAM after fine-tuning.

4 | DISCUSSION AND CONCLUSIONS

The current development of large foundation models, including SAM and MedSAM, offers potential promise to perform a unified solution for a large set of medical segmentation. Realizing that the importance of general purpose takes lower priority than stable accurate clinical performance, we performed comprehensive assessment and investigations of the large foundation model MedSAM and its light version LiteMedSAM to segment anatomical structures in pelvic MR images. We investigated the possible improvement (or the lack of) by using different bounding box prompts on MedSAM segmentations. More importantly, upon observing the general inferior performance on the specific task, we performed specialized fine-tuning and assessed its effectiveness.

Our results indicate that out-of-the-box MedSAM and LiteMedSAM exhibit suboptimal performance compared to state-of-the-art models regardless of the bounding box prompting schema used. We observed that the MedSAM models have a particularly difficult time segmenting objects that are non-convex or non-elliptical despite possessing relatively well-defined boundaries. This is likely because MedSAM was originally trained on

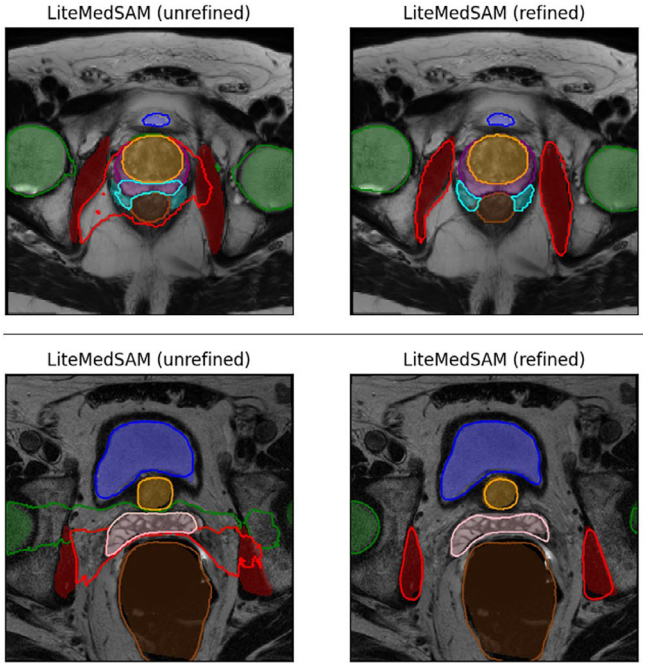

FIGURE 5 Visual comparison of LiteMedSAM before and after fine-tuning for the same two patients as shown in Figure 2.

a variety of modalities which included pathology images and dermoscopy, and a variety of segmentation tasks, including cellular and molecular, which may bias the segmentation decoder to contiguous convex shapes. We also found that MedSAM is challenged by segmenting multiple, disjointed objects.

Fine-tuning LiteMedSAM with a subset of our dataset yielded promising results, with the fine-tuned model performing comparably with nnU-Net, our benchmark model, especially for disjoint object segmentation. This suggests that MedSAM's foundation model can adapt effectively to specific task and anatomy with targeted training.

While our preliminary investigation shows that varying prompting using the current bounding box input offers only moderate to minimal improvement on the segmentation result, a more flexible and enabling scheme, such as wider shape atlas or customized masking,[2] may provide implicit guidance to the underlying task and could offer higher performance gain. Furthermore, multi-point prompting (a feature currently supported by SAM but not MedSAM) would allow the user to have more control over the demarcation of foreground/background which would also improve segmentation accuracy if implemented. These approaches offer an opportunity to combine a coarse task-specific training for mask prompt generation and the advantage of detail sensitivity from SAM/MedSAM's extensive training for feature encoding. In addition, the current work performed specialized refinement on the decoder portion of MedSAM to achieve comparable performance to the benchmark nnU-Net. It is expected that more sophisticated refinement, such as introducing modifiers to the deep layers of the encoder portion of MedSAM may further improve performance.

The model refinement test was only performed on LiteMedSAM due to the high memory requirement for training the complete MedSAM. Off-the-shelf, LiteMedSAM and MedSAM perform comparably, and prostate texture is relatively simple, so we believe that a LiteMedSAM is sufficient to capture the encoding power of MEDSAM with little compromise.

Isometric expansion of 5 pixels was used to explore relaxed bounding box prompting. As discussed earlier, refinement of the prompting may offer indirect injection of task-awareness into the decoding block in segmentation, but as a general theme, a good segmentation scheme should be reasonably robust against prompting.

The development of foundation models presents many potential clinical advantages. Clinics can test new tasks with minimal retraining, enabling faster adaptation to evolving clinical needs without extensive development efforts. Since these models leverage transfer learning and pre-training, they also require fewer annotated examples for fine-tuning on specific tasks, thus reducing the manual annotation burden.[18,19] However, before they can be implemented clinically, it is crucial to perform rigorous quality assurance and adjudication to appreciate its applicability and requirement. Despite their generalizability and efficiency advantages, they are not useful if they cannot provide accurate and robust results in a medical setting.

Overall, our study highlights the necessity and importance of specialized fine tuning to make large foundation models like MedSAM and LiteMedSAM to be clinically relevant and useful.

ACKNOWLEDGMENTS

The authors would like to acknowledge the supprot from UCLA Council of Research.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

REFERENCES

1. Bommasani R, Hudson DA, Adeli E, et al. On the Opportunities and Risks of Foundation Models. Published online July 12, 2022. doi:10.48550/arXiv.2108.07258
2. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. Published online April 5, 2023. doi:10.48550/arXiv.2304.02643
3. Mazurowski MA, Dong H, Gu H, Yang J, Konz N, Zhang Y. Segment anything model for medical image analysis: an experimental study. *Med Image Anal*. 2023;89:102918. doi:10.1016/j.media.2023.102918
4. Li K, Rajpurkar P, Adapting Segment Anything Models to Medical Imaging via Fine-Tuning without Domain Pretraining. In: ; 2024. Accessed April 3, 2024. https://openreview.net/forum?id=Fxi7pRmnYJ

5. Wu J, Ji W, Liu Y, et al. Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation. Published online December 28, 2023. doi:10.48550/arXiv.2304.12620
6. He S, Bao R, Li J, et al. Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets. Published online May 5, 2023. doi:10.48550/arXiv.2304.09324
7. Huang Y, Yang X, Liu L, et al. Segment anything model for medical images?. *Med Image Anal*. 2024;92:103061. doi:10.1016/j.media.2023.103061
8. Hu EJ, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models. Published online October 16, 2021. doi:10.48550/arXiv.2106.09685
9. Zhang K, Liu D, Customized Segment Anything Model for Medical Image Segmentation. Published online October 17, 2023. doi:10.48550/arXiv.2304.13785
10. Chen T, Zhu L, Ding C, et al. SAM Fails to Segment Anything?—SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More. Published online May 2, 2023. doi:10.48550/arXiv.2304.09148
11. Chen T, Zhu L, Ding C, et al. SAM-Adapter: adapting Segment Anything in Underperformed Scenes. In: *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE; 2023:3359-3367. doi:10.1109/ICCVW60793.2023.00361
12. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun*. 2024;15(1):654. doi:10.1038/s41467-024-44824-z
13. Li Y, Fu Y, Gayo I, et al. Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration. Published online 2022. doi:10.48550/ARXIV.2209.05160
14. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z
15. Isensee F, Jäger PF, Full PM, Vollmuth P, Maier-Hein KH. In: Crimi A, Bakas S, eds. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing; 2021:118-132. doi:10.1007/978-3-030-72087-2_11. nnU-Net for Brain Tumor Segmentation.
16. Pettit RW, Marlatt BB, Corr SJ, Havelka J, Rana A. nnU-net deep learning method for segmenting parenchyma and determining liver volume from computed tomography images. *Ann Surg Open*. 2022;3(2):e155. doi:10.1097/AS9.0000000000000155
17. Bhandary S, Kuhn D, Babaiee Z, et al. Investigation and benchmarking of U-Nets on prostate segmentation tasks. *Comput Med Imaging Graph*. 2023;107:102241. doi:10.1016/j.compmedimag.2023.102241
18. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *J Big Data*. 2016;3(1):9. doi:10.1186/s40537-016-0043-6
19. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. *BMC Med Imaging*. 2022;22(1):69. doi:10.1186/s12880-022-00793-7

How to cite this article: Nguyen E, Liu H, Ruan D. Necessity and impact of specialization of large foundation model for medical segmentation tasks. *Med Phys*. 2025;52:321–328. https://doi.org/10.1002/mp.17470