Personalizing Foundation Models for Cancer Imaging: A Study on Lymph Node Segmentation with SAM2 and MedSAM2

Accurate lymph node (LN) segmentation in abdominal CT is vital for cancer diagnosis and staging but remains a challenging task due to the small size of LNs, high inter-patient variability, low-contrast anatomical boundaries, and severe class imbalance. Traditional deep learning models often struggle in such settings, especially when annotated data is limited or the target structure occupies a minute portion of the image. With the emergence of large-scale vision foundation models like SAM2 and MedSAM2, there is now an opportunity to harness general-purpose and domain-specific pretrained architectures for precise medical image segmentation under data-scarce conditions. In this study, we systematically evaluate and compare **SAM2**[1] and **MedSAM2**[2] for the task of abdominal LN segmentation under three training regimes—zero-shot (no fine-tuning), few-shot (5 CTs), and big-shot (65 CTs)—using loss functions tailored for extreme pixel imbalance. We used the TCIA abdominal LN dataset[3] which has only ~13% of slices contain any lymph node annotations. Among these positive slices, over 97% have LN pixels occupying less than 1% of the image area, underscoring the substantial class and pixel-level imbalance that complicates automated segmentation.

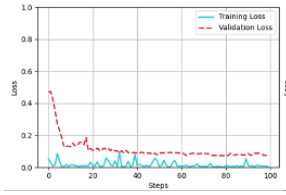

Figure 1 Paper loss

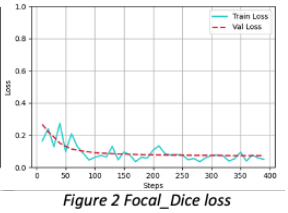

Figure 2 Focal_Dice loss

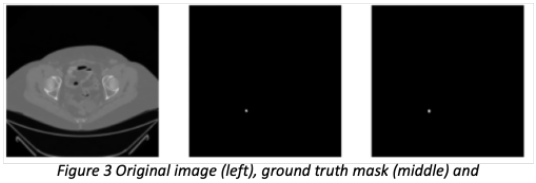

Figure 3 Original image (left), ground truth mask (middle) and MedSAM2's output (right)

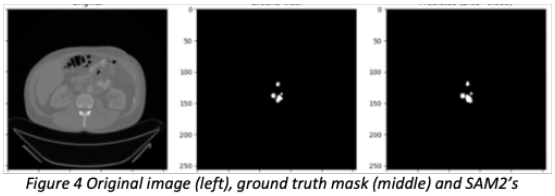

Figure 4 Original image (left), ground truth mask (middle) and SAM2's output (right)

For **SAM2**, we fine-tuned the decoder only and used single-slice bounding box prompts for each LN instance. We experimented with three loss combinations: BCE_Dice, Focal_Dice, and Focal Tversky. In contrast, **MedSAM2** was also fine-tuned at the decoder level but used default dual-slice bounding boxes for CT scans and supported two loss settings: a paper-defined loss (95.2% Focal + 4.8% Dice) and weighted BCE (WBCE). Performance was evaluated on 21 unseen validation scans (~1450 slices) using Dice and IoU metrics.

Table 1 MedSAM2 performance

Experiment	Validation Dice	Validation IOU	Training loss	Validation loss	Training time (hrs)
Zero-shot	0.751	0.656	---	---	---
Few-shot, WBCE	0.776252	0.672524	0.000799	0.307458	14.07
Big-shot, WBCE	0.8163	0.705583	0.003533	0.075518	20.91
Few-shot, Paper	0.819	0.717	0.006789	0.074725	22.262
Big-shot, Paper	0.835	0.745	0.044192	0.123963	28.32

Table 2 SAM2 performance

Experiment	Validation Dice	Validation IoU	Training loss	Validation loss	Training time (hrs)
Zero-shot	0.2432	0.1482	---	---	---
Few-shot, BCE_Dice (30%, 70%)	0.8092	0.6902	0.1537	0.1661	0.22
Big-shot, BCE_Dice	0.7906	0.6091	0.0876	0.2031	0.17
Few-shot, Focal_Dice	0.8204	0.6771	0.1384	0.1728	0.19
Big-shot, Focal_Dice	0.8307	0.7032	0.1763	0.1569	0.23
Few-shot, Focal_Tversky	0.8123	0.6858	0.2191	0.3069	0.21
Big-shot, Focal_Tversky	0.8234	0.6667	0.3385	0.3097	0.23

MedSAM2 achieved superior and **more stable performance**, with the paper loss yielding a peak Dice of 0.83 and smoother convergence (Figure 1). **SAM2**, while faster to train and still competitive in Dice (up to 0.83 using Focal Dice loss), showed higher fluctuations in training curves and less robustness under big-shot setups (Figure 2). Visual inspection of predicted masks (Figure 3–4) confirms that both models can delineate LN boundaries with high fidelity. Notably, **SAM2 demonstrated the ability to detect multiple LNs per slice**, while **MedSAM2's outputs appeared cleaner and more focused and generated results for instance segmentation**.

Performance summaries are presented in Tables 1 and 2. Despite SAM2's speed advantage, MedSAM2 is better suited for clinically sensitive tasks due to its consistency and medical-specific refinement and pretraining. Our findings underscore the potential of adapting foundation models with tailored fine-tuning strategies and appropriate losses for small-structure segmentation in medical imaging.

---

[1] https://github.com/facebookresearch/sam2

[2] https://github.com/SuperMedIntel/Medical-SAM2

[3] H. Roth *et al.*, "A new 2.5 D representation for lymph node detection in CT [Dataset]." The Cancer Imaging Archive, 2015. doi: 10.7937/K9/TCIA.2015.AQIIDCNM.