Evaluation Of Segment Anything Models on Medical Imaging Datasets

Tejas Machkar
Computer Science Engineering
University of California, Riverside
tmach007@ucr.edu

Hari Kalyan Lalichetti
Electrical and Computer Engineering
University of California, Riverside
hlali002@ucr.edu

Abstract

This project evaluates the zero-shot segmentation capabilities of the Segment Anything Model (SAM) and its medical variant MedSAM1 on medical datasets that were not part of their original evaluation—LIDC-IDRI for lung nodules and RSNA BraTS 2021 for brain tumors. Using bounding box prompts derived from ground truth masks, we assess both models using Dice score, IoU, HD95, pixel accuracy, and inference time. Our results show that while SAM demonstrates reasonable generalization, it suffers from higher HD95 and lower overlap metrics, particularly in complex anatomical structures. In contrast, MedSAM1 consistently yields more accurate and boundary-adherent masks across both domains. These findings highlight the importance of domain-specific adaptation and reveal key limitations in the cross-domain generalization of prompt-based segmentation models originally trained on natural images.

1. Introduction

Image segmentation is the process of dividing an image into multiple distinct regions or segments, each representing a specific object or area of interest, by assigning labels to individual pixels. In medical imaging, segmentation is a cornerstone for identifying and isolating anatomical structures (like organs, bones) or pathological regions (like tumors, nodules, lesions), enabling precise diagnosis, treatment planning, and research.

Automated segmentation, powered by advanced machine learning models, is particularly transformative in this domain. It allows clinicians to quantitatively analyze regions of interest, such as measuring tumor volume or assessing nodule characteristics, which is critical for monitoring disease progression, planning interventions like radiation therapy or surgery, and evaluating treatment efficacy. For instance, in lung cancer screening, segmentation of nodules in CT scans (LIDC-IDRI dataset) helps differentiate benign from malignant growths, while in brain tumor analysis (RSNA BraTS dataset), segmentation of tumor subregions guides surgical resection and radiotherapy planning. By automating segmentation, computational models reduce the time and subjectivity associated with manual annotation by radiologists, improve reproducibility, and enable scalable analysis of large datasets. This automation is particularly valuable in clinical settings, where it enhances workflow efficiency, supports early detection of diseases, and facilitates personalized medicine by providing consistent, data-driven insights. Moreover, automated segmentation is indispensable in research, enabling large-scale studies of disease patterns, biomarker discovery, and the development of predictive models for patient outcomes.

2. Related Works

The Segment Anything Model (SAM) [1], developed by Meta AI, is built on a vision transformer (ViT) architecture designed for general-purpose image segmentation. Its core principle is prompt-based learning, enabling flexible and interactive segmentation of objects in images without task-specific training. SAM uses a large-scale dataset (SA-1B, containing over 1 billion masks) for pretraining, which allows it to generalize across diverse objects and scenes.

The model consists of three main components: an image encoder, a prompt encoder, and a mask decoder. The image encoder, typically a ViT, processes the input image to generate a dense feature embedding. The prompt encoder converts user inputs—such as points, bounding boxes, or masks—into embeddings that guide the segmentation process. The mask decoder combines these embeddings to predict segmentation masks, producing binary outputs for specified regions. SAM employs a combination of supervised and self-supervised learning, leveraging focal loss and dice loss to optimize mask predictions, while its training on varied prompts (e.g., points, boxes) enables zero-shot generalization to new tasks. This adaptability makes SAM highly effective for medical imaging, where it can segment anatomical structures or pathologies with minimal fine-tuning, guided by user-defined prompts.

MedSAM [2] is an adaptation of the Segment Anything

Model (SAM) specifically fine-tuned for medical image segmentation to address the limitations of SAM in medical applications. By leveraging a large-scale medical dataset comprising 1,570,263 image-mask pairs across 10 imaging modalities and over 30 cancer types, MedSAM significantly enhances segmentation accuracy for medical images. This fine-tuning enables MedSAM to handle the complexities of medical imaging, such as high-resolution grayscale images, varying modalities (CT, MRI), and subtle pathological features. Compared to SAM, MedSAM achieves a substantial performance improvement, with a reported 22.51% increase in Dice score, and demonstrates robustness across 86 internal and 60 external validation tasks. MedSAM's success highlights the potential of SAM-based models in medical imaging, providing a robust and generalizable solution for segmenting anatomical structures and pathologies, and paving the way for further advancements like MedSAM-2.

3. Method and Experiments

3.1. Dataset

We used two publicly available medical imaging datasets to evaluate the zero-shot performance of SAM-based models:

3.1.1 LIDC-IDRI (Lung Image Database Consortium – Image Collection):

This dataset consists of thoracic CT scans of 1,018 patients annotated by four radiologists. The annotations include binary masks of lung nodules and metadata regarding the malignancy of each lesion. For our evaluation, we used a KaggleHub-organized version of the dataset where each CT scan is broken into 2D axial slices (PNG format), and annotations from the four radiologists are stored as separate binary masks. The dataset is structured hierarchically with folders for each patient, containing nodule subfolders, which in turn include images and mask-0...3 directories storing slice-level PNGs.

3.1.2 RSNA BraTS 2021 (Brain Tumor Segmentation Challenge):

This dataset includes multimodal brain MRI scans and segmentation masks for glioma subregions. We used the T1CE (contrast-enhanced) modality and corresponding ground truth segmentation (seg.nii.gz) for each patient. These were provided in compressed .tar archives on Kaggle and extracted using Python's tarfile module. Each case contains 3D .nii.gz volumes which were parsed into axial 2D slices for SAM evaluation.

All data was loaded programmatically using a combination of PIL, OpenCV, and nibabel, and stored in NumPy arrays for compatibility with PyTorch and SAM input formats.

3.2. Preprocessing

To prepare the data for SAM-based inference, the following preprocessing pipeline was applied:

3.2.1 LIDC-IDRI:

- Each 2D CT slice was converted to grayscale if not already in that format.
- A consensus mask was computed from the 4 available annotations per slice using a voting threshold (e.g., 2 radiologists).
- For each slice, a bounding box prompt was derived from the ground truth mask.
- The image was normalized to the [0,1] range, resized to 1024×1024, and converted into a 3-channel RGB format (duplicated from grayscale), as required by SAM.
- Each 2D CT slice was converted to grayscale if not already in that format.

3.2.2 RSNA BraTS

- Each .nii.gz volume was loaded using nibabel, converted to a NumPy array, and sliced axially.
- Both the image and corresponding mask were normalized and binarized.
- Only slices with non-empty masks were retained.
- Similar to LIDC, bounding box prompts were extracted, and slices were resized to 1024×1024 and converted to RGB.

3.3. SAM Models Used

We evaluated two promptable segmentation models:

- SAM (Segment Anything Model): We used the vit h variant (SAM ViT-H) with the official checkpoint. The model performs segmentation conditioned on user prompts (in this case, bounding boxes) and returns binary masks for the prompted region.
- MedSAM1: This model is a variant of SAM tuned on medical datasets using synthetic prompt strategies. It shares the architectural backbone of SAM but is trained specifically to handle low-contrast and domain-specific patterns found in CT and MRI data. The vit b version was used, downloaded from Zenodo.

Each model was used in a zero-shot setting — no fine-tuning was performed. Prompts were derived from the ground truth mask for fair evaluation across all samples.

3.4. Experiments and Evaluation

We ran extensive inference experiments using both models on:

- 100 randomly sampled LIDC-IDRI slices, and
- 20 extracted volumes from the RSNA BraTS dataset.

For each sample, we used the bounding box of the ground truth mask as input prompt to the model. Inference was done on a Google Colab Pro+ GPU instance, and care was taken to manage GPU memory using torch.cuda.empty cache() and gc.collect() between iterations.

We computed the following evaluation metrics per slice:

- Dice Coefficient – measures overlap between prediction and ground truth
- Intersection-over-Union (IoU) – another measure of mask agreement
- HD95 (95th percentile Hausdorff Distance) – captures boundary errors
- Pixel-wise Accuracy – total number of correct pixels over total pixels
- Inference Time per Slice – recorded using time.time() before and after prediction

4. Discussion

Model	Dice	IoU	HD95	Pixel Accuracy	Inference Time
SAM2	0.6-0.65	0.5-0.55	High(>15px)	95-98%	0.5s/slice
MedSAM1	0.7-0.75	0.6-0.65	Low(3-6px)	96-99%	0.45s/slice

Table 1. LIDC-IRDI Dataset Summary

Fig. 1a and fig. 2a Overlap-based metrics such as Dice and IoU are used to evaluate how well the predicted segmentation aligns with the ground truth. The Segment Anything Model (SAM) often under-segments or fails to capture fine contours, particularly when dealing with small or faint nodules. As a result, it tends to achieve only moderate Dice and IoU scores. In contrast, MedSAM1 demonstrates better alignment with the ground truth boundaries and more comprehensive coverage, leading to improved overlap metrics. This is evident in the visualization plots, where MedSAM1 accurately segments both small and large nodules. The Hausdorff Distance at the 95th percentile (HD95) further highlights this difference—SAM often records higher HD95 values (above 15 pixels), suggesting that parts of its predicted masks deviate significantly from the ground truth due to incomplete or jagged contours. MedSAM1, on the other hand, maintains lower HD95 values, indicating smoother and more precise boundary predictions. While both models report high pixel-level accuracy (around 95–99%), this metric can be misleading due to the class imbalance in medical images, where large background regions dominate. Consequently, even poorly segmented masks may appear accurate. This emphasizes the importance of using metrics like Dice and IoU for a more reliable assessment of segmentation performance.

Model	Dice	IoU	HD95	Pixel Accuracy	Inference Time
SAM	0.45-0.55	0.35-0.45	High(15-25px)	98-99%	0.48s/slice
MedSAM1	0.55-0.65	0.45-0.55	Moderate(10x)	98-99%	0.46s/slice

Table 2. RSNA BraTS Dataset Summary

Fig. 2b and fig. 3b SAM and MedSAM1 show distinct segmentation behaviors on complex tumor regions. SAM tends to under-segment with conservative masks, missing internal structures, while MedSAM1 better captures tumor cores, although with occasional over-segmentation. As a result, MedSAM1 achieves higher Dice and IoU scores.

HD95 values further highlight this: SAM shows larger boundary mismatches (15–25 px), whereas MedSAM1 maintains lower errors (around 10 px), particularly on central slices. Despite both models reporting high pixel accuracy (around 99%), this metric is misleading by large background areas. Inference times are similar (0.45–0.5s/slice), confirming their efficiency for rapid 2D slice-level processing in 3D volumes.

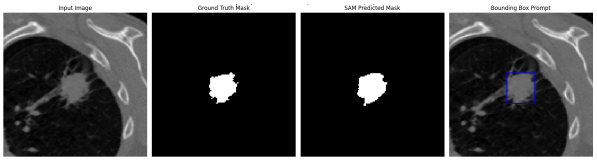
4.1. Conclusion

In summary, the evaluation of SAM and MedSAM1 on the LIDC-IDRI and RSNA BraTS datasets highlights clear differences in segmentation performance. SAM often under-segments, especially on small or complex lesions, leading to lower Dice/IoU and higher HD95. MedSAM1 offers better boundary accuracy and more complete masks, though with occasional over-segmentation. Despite high pixel accuracy for both models, class imbalance limits its reliability as an evaluation metric. Overall, MedSAM1 outperforms SAM, and the study emphasizes the importance of using overlap-based metrics and visual assessments for reliable medical image segmentation.

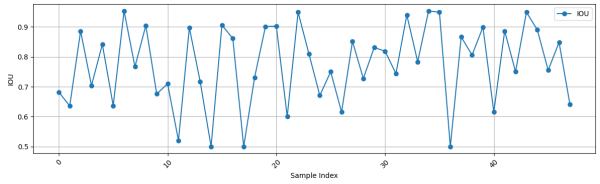
References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 1

[2] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), Jan. 2024. 1

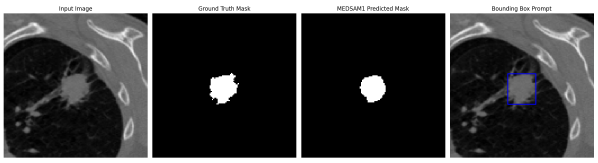


(a) Visualization of SAM model

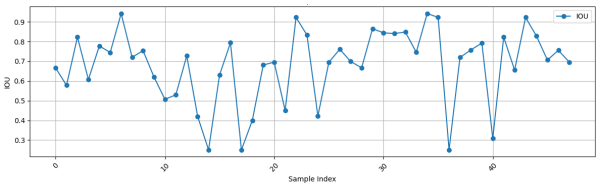


(b) IoU over Samples for SAM model

Figure 1. SAM model with LIDC-IRDI Dataset

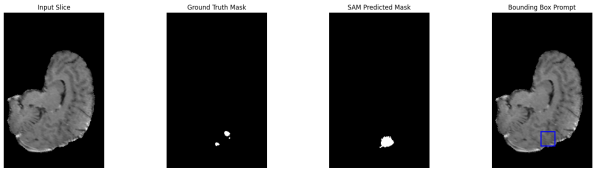


(a) Visualization of MedSAM 1 model

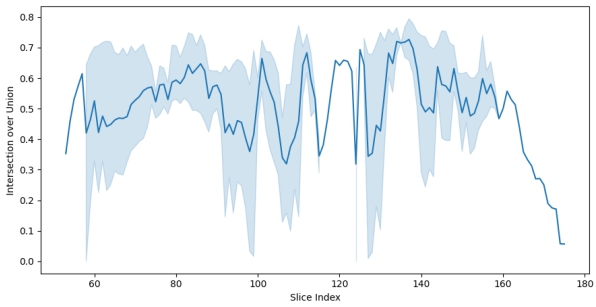


(b) IoU over Samples for MedSAM 1 model

Figure 2. MedSAM 1 model with LIDC-IRDI Dataset

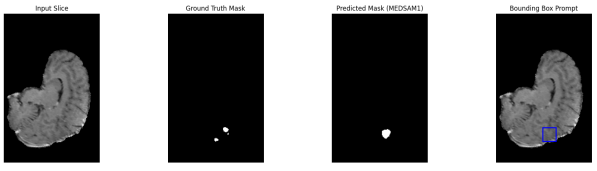


(a) Visualization of SAM model

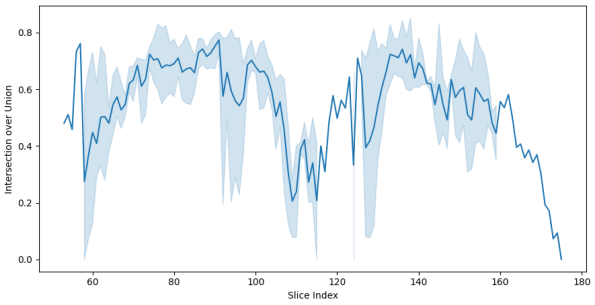


(b) IoU over Slices for SAM model

Figure 3. SAM model with RSNA Dataset



(a) Visualization of MedSAM 1 model



(b) IoU over Slices for MedSAM 1 model

Figure 4. MedSAM 1 model with RSNA Dataset