Benchmarking Segment Anything Model on Medical Segmentation Decathlon

Cristina-Diana Savin
Faculty of Mathematics and Computer Science
University of Bucharest
Bucharest, Romania
cristina-diana.savin@s.unibuc.ro

***Abstract*—We benchmark the Segment Anything Model (SAM) and its medical variant, MedSAM, on the Medical Segmentation Decathlon (MSD). Using bounding box prompts, we evaluate segmentation performance across ten medical tasks and multiple modalities. Results show that MedSAM excels with precise prompts on specific organs, while SAM is more robust to varying input conditions. We highlight key differences and provide recommendations for applying prompt-based segmentation in medical imaging.**

***Index Terms*—Medical image segmentation, Segment Anything Model, MedSAM, Medical Segmentation Decathlon, prompt-based segmentation, bounding box prompts, evaluation metrics.**

I. Introduction

Accurate segmentation in medical imaging is vital for diagnosis, treatment planning, and clinical workflows. However, building robust models typically requires large volumes of expert-labeled data—an expensive and domain-specific challenge.

The Segment Anything Model (SAM), developed by Meta AI, enables zero-shot segmentation using prompts like bounding boxes and has demonstrated impressive generalization across natural images. MedSAM extends this idea by fine-tuning SAM on medical datasets, aiming to bridge the domain gap between natural and medical imagery.

In this study, we evaluate SAM and MedSAM on the Medical Segmentation Decathlon (MSD), a benchmark spanning 10 tasks across multiple organs and imaging modalities. We assess segmentation performance under varying bounding box sizes and analyze model robustness, task difficulty, and metric trends.

Our goal is to understand when SAM or MedSAM is more effective and to derive practical recommendations for prompt-based segmentation in medical applications.

II. Dataset and Tasks Overview

We use the Medical Segmentation Decathlon (MSD) as the benchmark dataset for evaluating SAM and MedSAM. MSD consists of ten segmentation tasks across a variety of anatomical structures, imaging modalities, and labeling challenges. Each task contains 3D volumetric scans in NIfTI format, with corresponding ground truth segmentations.

The dataset covers three main imaging modalities: Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and FLAIR. Depending on the task, there may be one to three labeled structures per volume. Tasks vary significantly in complexity, from large organs (e.g., liver, spleen) to small or irregular structures (e.g., hippocampus, vessels).

TABLE I
MSD Tasks Overview

Task	Modality	Target Organ(s)	Labels
Task01	MRI	Brain Tumour	3
Task02	CT	Heart	1
Task03	CT	Liver	2
Task04	MRI	Hippocampus	2
Task05	MRI	Prostate	2
Task06	MRI	Lung	1
Task07	CT	Pancreas	2
Task08	CT	Hepatic Vessel	1
Task09	CT	Spleen	1
Task10	CT	Colon	1

For this study, each 3D scan is processed slice-by-slice in 2D. Preprocessing includes modality filtering, label thresholding, and optional resizing to standardize input resolution across tasks. The bounding box used to prompt segmentation is derived from the minimal enclosing rectangle of each labeled region, then optionally enlarged by a fixed pixel margin.

III. Methodology

A. Bounding Box Prompting

We evaluate the promptable performance of SAM and MedSAM using bounding box inputs. For each ground truth mask, we compute the tightest bounding box that encloses the labeled region. To simulate different levels of contextual information, we enlarge this box by fixed pixel margins ranging from 0 to 20 pixels in each direction. This approach allows us to assess model sensitivity to prompt tightness and background noise.

B. Input Preparation

All images and labels are extracted as 2D slices from the original 3D NIfTI volumes. Only slices containing at least one non-zero label are included. For multi-label tasks, all foreground labels are merged into a single binary mask. Input images are normalized, resized when necessary, and processed per-slice. Both SAM and MedSAM receive the same input slice and bounding box prompt for fair comparison.

C. Evaluation Metrics

We use two primary metrics to evaluate segmentation quality:

Dice Similarity Coefficient (DSC) measures overlap between the predicted and ground truth masks. It is defined as:

$$\text{DSC} = \frac{2|P \cap G|}{|P| + |G|}$$

where P is the predicted mask and G is the ground truth.

Normalized Surface Dice (NSD) evaluates boundary alignment by computing the fraction of the predicted surface within a tolerance distance of the ground truth surface. NSD is more sensitive to boundary accuracy than overall overlap.

These two metrics provide different perspectives: DSC shows how well the predicted region overlaps with the ground truth, while NSD focuses on how well the boundaries match. We report both metrics on a per-slice and per-task basis. Inference time and basic pixel statistics are also recorded to support additional comparisons.

IV. Experiment Setup

A. Model Versions

We used the official SAM implementation `sam_vit_b` with checkpoint `sam_vit_b_01ec64.pth`. MedSAM was loaded from `medsam_vit_b.pth`, a fine-tuned variant adapted for medical imaging. Both models were evaluated using the same prompt interface through `SamPredictor` to ensure consistency.

B. Prompts and Parameters

No manual tuning or additional user input was used. The only prompt was an automatically computed bounding box per labeled region, optionally enlarged by 0, 3, 5, 10, 15, or 20 pixels. No post-processing was applied to predictions. For each configuration, the model was applied to all eligible slices in a given task to compute per-slice metrics.

V. Results

A. Quantitative Performance by Bounding Box Size

Table II shows average Dice and NSD scores for SAM and MedSAM across different bounding box enlargements. Smaller boxes (0 to 5 pixels) consistently yield better performance, especially for MedSAM.

TABLE II
AVERAGE DICE AND NSD BY BOUNDING BOX SIZE (ALL TASKS)

BBox	Dice (SAM)	NSD (SAM)	Dice (MedSAM)	NSD (MedSAM)
0	0.6786	0.4570	0.5808	0.3051
3	0.7024	0.3977	0.6671	0.3464
5	0.6663	0.3422	0.6593	0.3284
10	0.5957	0.2643	0.6026	0.2413
15	0.5332	0.2031	0.5272	0.1612
20	0.4751	0.1593	0.4572	0.1087

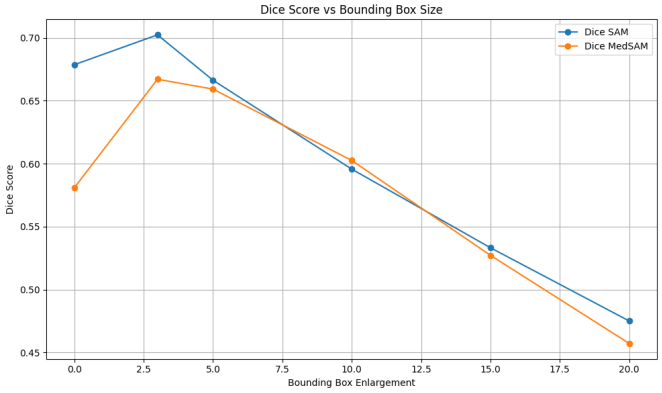

Fig. 1. Dice vs Bounding Box Enlargement for SAM and MedSAM.

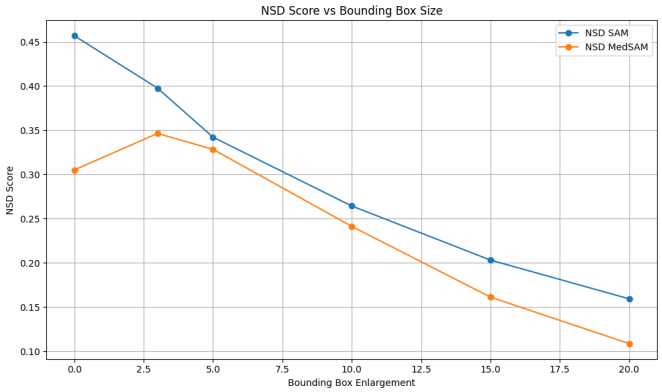

Fig. 2. NSD vs Bounding Box Enlargement for SAM and MedSAM.

B. Per-Task Performance

Figure 3 shows the Dice scores for each MSD task. SAM performs better on tasks like prostate, pancreas, and lung, while MedSAM excels in spleen and hippocampus.

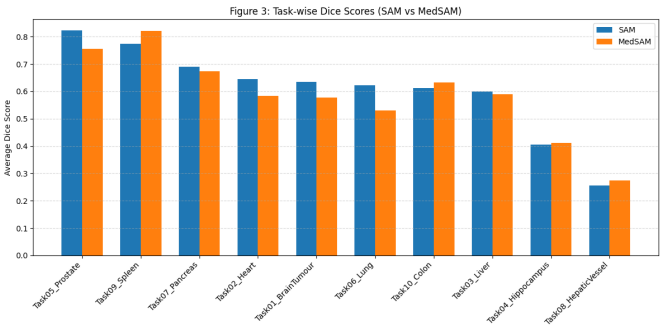

Fig. 3. Per-task Dice comparison between SAM and MedSAM.

C. Task Sensitivity and Model Differences

We identified tasks with the highest and lowest performance. Table III summarizes key findings.

TABLE III
EXTREME PERFORMANCE CASES AND NOTABLE TRENDS

Insight	Task	BBox
Best Dice (SAM)	Spleen	0
Best Dice (MedSAM)	Spleen	5
Worst Dice (SAM)	Hippocampus	20
Worst Dice (MedSAM)	HepaticVessel	0
Highest NSD Gain (MedSAM-SAM)	Hippocampus	15

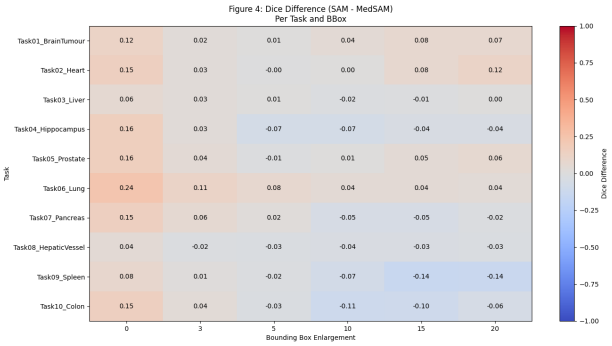

Fig. 4. Heatmap of Dice differences (SAM - MedSAM) across tasks and bounding box sizes.

D. Summary Metrics

Table IV aggregates the overall performance across all tasks and prompts.

TABLE IV
MEAN ± STD DICE AND NSD ACROSS ALL TASKS

Metric	SAM	MedSAM
Dice	0.626 ± 0.085	0.599 ± 0.092
NSD	0.304 ± 0.109	0.248 ± 0.104

E. Inference Time

Both models complete inference in under 0.41 seconds per slice. Minor variation is observed across tasks due to differences in slice size and organ complexity. No significant performance gap was measured between SAM and MedSAM in terms of runtime.

F. Visual Example: Brain Tumour Segmentation

To illustrate model behavior, Fig. 5 shows the segmentation outputs for a slice from Task01_BrainTumour using SAM and MedSAM. Despite both models achieving similar Dice scores (0.786 for SAM vs 0.767 for MedSAM), their boundary alignment and shape predictions differ slightly, as reflected in the NSD values.

VI. INTERACTIVE STREAMLIT APPLICATION

To support visual exploration and qualitative comparison, we developed a lightweight Streamlit application. The app allows users to upload a medical image slice, define a bounding box interactively via mouse click, and view segmentation results from both SAM and MedSAM side by side.

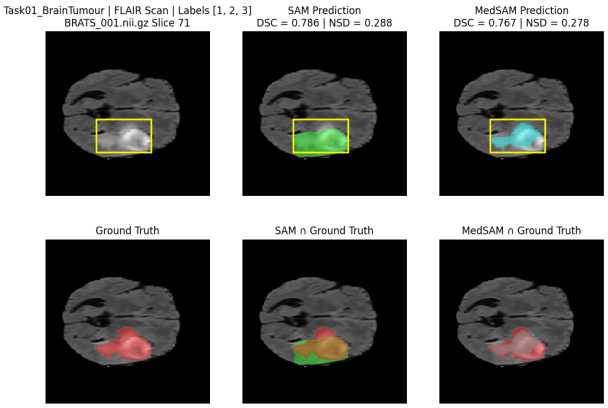

Fig. 5. Qualitative comparison of SAM and MedSAM on a brain tumour slice (Task01). Top row: input with bounding box, SAM and MedSAM predictions. Bottom row: ground truth and overlap with predictions.

This tool is intended for demonstration and educational purposes. It provides an intuitive way to explore how bounding box placement affects outputs.

VII. KEY INSIGHTS

A. Bounding Box Size Matters

Smaller bounding boxes consistently improved performance. Dice and NSD scores were highest when the prompt tightly enclosed the target (bbox = 0, 3, or 5). Enlarging the bounding box beyond 10 pixels degraded performance, especially for MedSAM, likely due to added background noise.

B. MedSAM Benefits from Focused Prompts

MedSAM achieved its best Dice on the spleen (0.821) and hippocampus tasks when bounding boxes were small or moderately expanded. However, its performance dropped more steeply than SAM as the bounding box size increased (see Fig. 1).

C. SAM Is More Robust to Noisy Prompts

SAM showed greater consistency across bounding box sizes. In tasks like lung and prostate segmentation, SAM outperformed MedSAM at larger box sizes, indicating better resilience to context noise.

D. Dice and NSD Diverge on Certain Tasks

In some cases (e.g., hippocampus), MedSAM achieved better NSD than SAM despite similar or lower Dice scores. This suggests MedSAM may generate more accurate boundaries even if the overall overlap is not maximal.

E. Task Difficulty Varies Widely

Tasks involving thin or irregular structures—like hepatic vessels or hippocampus—produced lower Dice scores for both models (see Table III). This suggests a need for task-specific post-processing or model tuning.

F. Model Selection Should Be Task-Aware

A hybrid strategy is recommended: use MedSAM when the bounding box is tight and structure is well-localized (e.g., spleen, hippocampus, heart), and SAM when dealing with noisy prompts or larger, complex regions (e.g., lung, prostate).

VIII. Recommendations

Based on our benchmarking results, we provide the following practical recommendations for using prompt-based segmentation in medical imaging:

A. Bounding Box Size

A small enlargement of 3–5 pixels around the ground truth region provides the best trade-off between context and noise. Avoid large expansions (bigger than 15 pixels), as they significantly reduce segmentation quality, especially for MedSAM.

B. Model Choice by Task

MedSAM performs best on well-localized, compact structures such as the spleen, hippocampus, and heart—especially when the bounding box is precise. For these tasks, MedSAM is recommended.

SAM is more stable across a range of bounding box sizes and is preferred for tasks involving:

- Noisy or poorly localized prompts;
- Large or irregular structures (e.g., lung, prostate);
- Use cases requiring prompt robustness.

C. Hybrid Strategy

A hybrid deployment strategy is advised. Use task-specific heuristics or prompt quality estimators to select between SAM and MedSAM dynamically. This balances MedSAM's high precision with SAM's general robustness.

D. Future Use and Extensions

Further improvements could be achieved by:

- Using the predicted mask from a previous slice as a spatial prior for the current slice. This could improve continuity and stability across adjacent slices in 3D volumes.
- Training MedSAM variants on smaller structures.
- Exploring multi-modal fusion prompts.

IX. Limitations

This study focuses on 2D slice-wise evaluation, which does not fully capture the 3D context available in medical volumes. While sufficient for prompt-based segmentation benchmarking, it may underestimate the models' potential in volumetric settings.

We did not apply any post-processing, such as connected component filtering or smoothing, which could improve results, especially for fragmented predictions.

Modality-specific effects (e.g., MRI vs CT) were not isolated, and no modality-adaptive prompting or fusion was used.

Finally, our evaluation does not include expert clinical review or comparison with fully supervised task-specific baselines, which may outperform both SAM and MedSAM in ideal conditions.

X. Conclusion

We benchmarked the Segment Anything Model (SAM) and its medical variant, MedSAM, on the Medical Segmentation Decathlon (MSD), using bounding box prompts to evaluate prompt sensitivity and task-specific performance.

Our results show that bounding box size strongly affects segmentation quality. MedSAM achieves higher accuracy on certain tasks when prompts are tight and well-localized, while SAM is more robust to noisy or enlarged prompts.

Dice and NSD trends reveal trade-offs between overlap and boundary accuracy, and no single model outperforms the other universally.

We recommend using MedSAM for precise, localized structures like the spleen and hippocampus, and SAM for broader or noisier tasks like lung or prostate segmentation. A hybrid strategy combining both models based on task or prompt conditions offers the best performance and flexibility.

This work highlights the potential of prompt-based segmentation in medical imaging and suggests practical directions for deployment in real-world clinical pipelines.

Acknowledgment

This work was conducted as part of a research project within the Faculty of Mathematics and Computer Science at the University of Bucharest. The author would like to thank Teaching Assistant Ciprian Ceaușescu for his guidance and support.

The project made use of open-source tools including the Segment Anything Model (Meta AI), MedSAM, Streamlit, and the Medical Segmentation Decathlon dataset. The author also acknowledges the contributors of these resources for enabling reproducible research.

References

[1] M. Kirillov, E. Mintun, N. Ravi, et al., "Segment Anything," *arXiv preprint arXiv:2304.02643*, 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[2] Z. Ma, H. Chen, Y. Wang, et al., "Segment Anything in Medical Images," *Nature Communications*, vol. 15, no. 1, 2024. [Online]. Available: https://www.nature.com/articles/s41467-024-44824-z

[3] Medical Segmentation Decathlon. [Online]. Available: http://medicaldecathlon.com/

[4] T. Heimann, et al., "The Medical Segmentation Decathlon," *arXiv preprint arXiv:1902.09063*, 2019. [Online]. Available: https://arxiv.org/abs/1902.09063

[5] Segment Anything GitHub Repository. Meta AI. [Online]. Available: https://github.com/facebookresearch/segment-anything

[6] MedSAM GitHub Repository. Bo Wang Lab. [Online]. Available: https://github.com/bowang-lab/MedSAM

[7] Streamlit Documentation. [Online]. Available: https://streamlit.io/