MedSAM2: Segment Anything in 3D Medical Images and Videos

Jun Ma*, Zongxin Yang*, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang†

Abstract

Medical image and video segmentation is a critical task for precision medicine, which has witnessed considerable progress in developing task or modality-specific and generalist models for 2D images. However, there have been limited studies on building general-purpose models for 3D images and videos with comprehensive user studies. Here, we present MedSAM2, a promptable segmentation foundation model for 3D image and video segmentation. The model is developed by fine-tuning the Segment Anything Model 2 on a large medical dataset with over 455,000 3D image-mask pairs and 76,000 frames, outperforming previous models across a wide range of organs, lesions, and imaging modalities. Furthermore, we implement a human-in-the-loop pipeline to facilitate the creation of large-scale datasets resulting in, to the best of our knowledge, the most extensive user study to date, involving the annotation of 5,000 CT lesions, 3,984 liver MRI lesions, and 251,550 echocardiogram video frames, demonstrating that MedSAM2 can reduce manual costs by more than 85%. MedSAM2 is also integrated into widely used platforms with user-friendly interfaces for local and cloud deployment, making it a practical tool for supporting efficient, scalable, and high-quality segmentation in both research and healthcare environments.

arXiv:2504.03600v1 [eess.IV] 4 Apr 2025

INTRODUCTION

Medical image segmentation plays a pivotal role in numerous clinical applications, including anatomical structure analysis [1], disease diagnosis [2], surgery planning, and treatment monitoring [3]. By delineating the boundaries of organs, lesions, and other relevant anatomies, segmentation algorithms provide clinicians with crucial information for precise disease analysis. Over the past decade, deep learning-based methods have revolutionized this field, delivering unprecedented performance on various segmentation tasks and benchmarks. For example, DeepLab [4] [5] has achieved human-level performance in left ventricle segmentation from echocardiography for ejection fraction assessment [1], which has proven to save time for both sonographers and cardiologists via blinding and randomization clinical trial [6]. U-Net [7] has been employed for accurate cell detection and segmentation in light microscopy images [8] and 3D nnU-Net [9] has been widely used in various anatomy and lesion segmentation, such as heart chamber segmentation in Magnetic Resonance Imaging (MRI) scans [2], pancreas cancer and abdominal organ segmentation in Computed Tomograph (CT) scans [3] [10], and whole-body lesion segmentation in Positron Emission Tomography (PET) scans [11].

Driven by advanced network architectures [12] and large-scale datasets [13], recent trends in segmentation present a paradigm shift from specialist models tailored for specific tasks to generalist or foundation models capable of performing segmentation without extensive task-specific model development [14]–[16]. One prominent example is the Segment Anything Model (SAM) [13], a pioneer segmentation foundation model in computer vision that has shown remarkable generalization ability across a wide range of two-dimensional (2D) natural image segmentation tasks. However, due to the substantial domain gap, its performance remains suboptimal in medical images [17] [18]. Despite these limitations, SAM can be effectively adapted to the medical domain through transfer learning. For instance, models such as MedSAM [19] and SAM-Med [20] [21] have demonstrated strong capabilities in segmenting various organs and abnormalities across diverse medical imaging modalities by fine-tuning SAM on large-scale medical datasets.

- *Jun Ma is with AI Collaborative Centre, University Health Network; Vector Institute, Toronto, Canada (∗ Equal Contribution).*
- *Zongxin Yang is with Department of Biomedical Informatics, Harvard Medical School, Harvard University, Boston, USA (∗ Equal Contribution).*
- *Sumin Kim is with Peter Munk Cardiac Centre, University Health Network; Department of Computer Science, University of Toronto; Vector Institute, Toronto, Canada.*
- *Bihui Chen is with Peter Munk Cardiac Centre, University Health Network; Department of Computer Science, University of Toronto; Vector Institute, Toronto, Canada.*
- *Mohammed Baharoon is with Department of Biomedical Informatics, Harvard Medical School, Harvard University, Boston, USA. Part of this work was done at the University of Toronto, Toronto, Canada.*
- *Adibvafa Fallahpour is with Peter Munk Cardiac Centre, University Health Network; Department of Computer Science, University of Toronto; Vector Institute, Toronto, Canada.*
- *Reza Asakereh participated in this project when he was with Peter Munk Cardiac Centre, University Health Network, Toronto, Canada.*
- *Hongwei Lyu is with Peter Munk Cardiac Centre, University Health Network, Toronto, Canada.*
- *Bo Wang is with Peter Munk Cardiac Centre and AI Hub, University Health Network; Department of Laboratory Medicine and Pathobiology and Department of Computer Science, University of Toronto; Vector Institute, Toronto, Canada(†Corresponding Author). E-mail: bowang@vectorinstitute.ai*

Despite the potential of these foundation models, their application to medical imaging is still limited and faces three main limitations. First, most medical image segmentation foundation models [19] [20] are primarily designed for 2D image data and may not capture the three-dimensional (3D) spatial relationships or temporal information in volumetric and video medical data. Second, although some studies have extended SAM to 3D image segmentation using 3D image encoders [21] and adapters [22]–[24] or developed interactive 3D segmentation models [25]–[27] to incorporate manual corrections, there is still a lack of general models to segment both 3D images and videos, which are frequently necessary in real-world clinical workflows. The state-of-the-art video segmentation model, SAM2 [28], has shown great potential to fill this gap [29]–[32], but adaption on large-scale datasets has been underexplored. Finally, large-scale validation of these models in practical image-labeling scenarios remains notably absent, leaving important questions about their scalability and utility in facilitating high-throughput medical image annotation tasks.

In this work, we address these limitations by presenting MedSAM2, a general model for 3D medical image and video segmentation. Specifically, we first curate a large-scale dataset consisting of more than 455,000 3D image–mask pairs and 76,000 annotated video frames, spanning multiple organs, pathologies, and imaging protocols for model development. Then, we build MedSAM2 by modifying and fine-tuning SAM2 on the large dataset. Extensive experiments show that MedSAM2 is capable of handling both volumetric medical scans and successive video frames, enabling versatile segmentation across diverse medical data. Furthermore, we conduct three user studies to demonstrate that MedSAM2 substantially facilitates annotation workflows for high-throughput and efficient segmentation, substantially reducing the time and effort required for creating large-scale medical datasets in various imaging modalities. MedSAM2 has the potential to transform clinical workflows by enabling more efficient diagnostic processes, treatment planning, and longitudinal monitoring across cardiology, oncology, and surgical specialties, where precise 3D organ and lesion segmentation is critical but traditionally time-consuming.

RESULTS

Dataset and model architecture

A large amount of training data is the foundation for developing generalist segmentation models. We assembled a large-scale and diverse 3D medical image and video dataset based on public datasets, including various normal anatomical structures and pathologies from various medical imaging modalities (Fig 1a, Methods, Supplementary Table 1). In particular, we collected 363,161, 14,818, and 77,154 3D image-mask pairs for CT, PET, and MRI modalities, respectively. In addition, we curated 19,232 and 56,462 annotated frames for ultrasound and endoscopy, respectively.

The pre-trained SAM2 model [28] has provided a strong backbone for general feature representations, which was trained on 256 A100 GPUs. To reuse the pre-trained model weights and avoid prohibitive computing costs, MedSAM2 adopted the SAM2 network architecture, including an image encoder, a memory attention module, a prompt encoder, and a mask decoder (Fig 1b). The image encoder extracts multi-scale features from each 2D slice or video frame using the hierarchical vision transformer [33] (Hiera), which achieves faster and more accurate performance than the naÏ've vision transformer [12] in SAM. The memory attention module employs transformer blocks with self-attention and cross-attention mechanisms to condition current frame features on previous frames' predictions through a streaming memory bank. The prompt encoders convert various user interactions (*i.e.*, points, bounding boxes, and masks) to embedding. We used bounding boxes as the main prompt because they are less ambiguous in specifying the segmentation target, making them suitable for most organs and lesions. Specifically, for 3D images, we applied the bounding box prompt on the middle slice and propagated the segmentation mask bidirectionally toward both ends of the volume data. Finally, the mask decoder incorporates memory-conditioned features and prompt embeddings to produce accurate segmentation masks.

Existing studies have demonstrated that fine-tuning all parts of the model yields better performance than only fine-tuning parts of the model, such as the image encoder, the mask decoder, and the prompt encoder [34], [35]. For MedSAM2, we employ a comprehensive full-model fine-tuning approach using the lightweight SAM2.1-Tiny variant, which achieved competitive performance with fewer parameters compared to larger variants. During fine-tuning, we applied lower learning rates for the image encoder to preserve pre-trained feature extraction capabilities and higher learning rates for other model parts. We carefully balanced our training data with different sampling rates across 3D images and videos to ensure optimal performance across diverse modalities (Methods).

Performance on various 3D medical image and video segmentation tasks

We first evaluated the trained model on the holdout 3D test set, which contains 40 segmentation tasks from different cohorts across a wide range of organs and lesions in CT, MRI, and PET scans. We also compared the latest SAM2.1 models with different sizes (tiny, small, base, and large) [28] and the current state-of-the-art (SOTA) bounding box-based segmentation foundation model (EfficientMedSAM-Top1) [36], which is the winning solution in the CVPR 2024 Efficient MedSAMs competition [37]. All models were initialized with a bounding box prompt on the middle slice of the segmentation target. Each model first generated a 2D mask at the middle slice and then propagated it bidirectionally to create the full 3D segmentation.

Fig. 2a shows the quantitative results on the 3D testing set (Supplementary Table 2-3 and Fig. 1). The SAM2.1 models exhibit similar performance across all categories, with no significant differences in median DSC scores ($p-$value