SAM-Med3D: Towards General-purpose Segmentation Models for Volumetric Medical Images

Haoyu Wang[1,2], Sizheng Guo[1], Jin Ye[1], Zhongying Deng[1], Junlong Cheng[1], Tianbin Li[1], Jianpin Chen[1], Yanzhou Su[1], Ziyan Huang[1,2], Yiqing Shen[1], Bin Fu[3], Shaoting Zhang[1], Junjun He[1]*, and Yu Qiao[1]*

[1] Shanghai Jiao Tong University
[2] Shanghai Artificial Intelligence Laboratory
[3] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
`hejunjun@pjlab.org.cn`

Abstract. Existing volumetric medical image segmentation models are typically task-specific, excelling at specific target but struggling to generalize across anatomical structures or modalities. This limitation restricts their broader clinical use. In this paper, we introduce SAM-Med3D for general-purpose segmentation on volumetric medical images. Given only a few 3D prompt points, SAM-Med3D can accurately segment diverse anatomical structures and lesions across various modalities. To achieve this, we gather and process a large-scale 3D medical image dataset, SA-Med3D-140K, from a blend of public sources and licensed private datasets. This dataset includes 22K 3D images and 143K corresponding 3D masks. Then SAM-Med3D, a promptable segmentation model characterized by the fully learnable 3D structure, is trained on this dataset using a two-stage procedure and exhibits impressive performance on both seen and unseen segmentation targets. We comprehensively evaluate SAM-Med3D on 16 datasets covering diverse medical scenarios, including different anatomical structures, modalities, targets, and zero-shot transferability to new/unseen tasks. The evaluation shows the efficiency and efficacy of SAM-Med3D, as well as its promising application to diverse downstream tasks as a pre-trained model. Our approach demonstrates that substantial medical resources can be utilized to develop a general-purpose medical AI for various potential applications. Our dataset, code, and models are available at: https://github.com/uni-medical/SAM-Med3D.
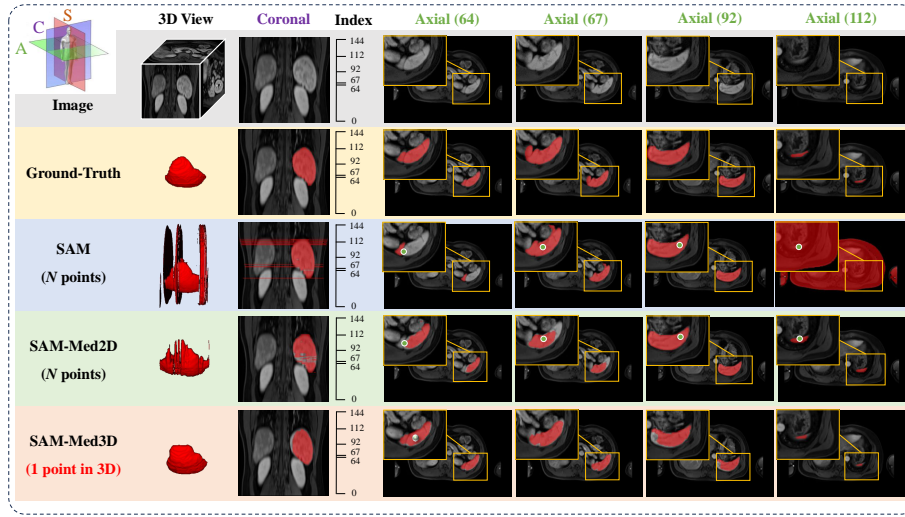
Keywords: General-purpose segmentation · Volumetric medical images

1 Introduction

Medical image analysis has become an indispensable cornerstone of modern healthcare, aiding disease diagnosis, treatment planning, education, and medical

---

* Corresponding authors.

Fig. 1: Illustration of SAM [18], fine-tuned SAM (SAM-Med2D [6]), and our SAM-Med3D on 3D Volumetric Medical Images. Both SAM and SAM-Med2D take N prompt points (one for each slice) whereas SAM-Med3D uses a single prompt point for the entire 3D volume. Here, N corresponds to the number of slices containing the target object. The top-left corner provides a schematic of the Axial, Coronal, and Sagittal views. For a given 3D input, we visualize the 3D, coronal, and multiple axial views. The numbers in brackets indicate the index of each axial slice.

research [15, 32, 34]. One of the major challenges in this realm is the precise segmentation of volumetric medical images [21]. Although numerous methods have demonstrated commendable effectiveness across a spectrum of targets [22, 24, 39], mainstream segmentation models tend to specialize in particular organs or lesions. This tendency is attributed to the inherent characteristics of public volumetric medical image datasets: most datasets in medical imaging are collected for specific tasks such as CT images labeled only for organs segmentation [36], or PET & CT images with tumor labels for lesion detection [9]. While models trained on such specialized datasets excel in their respective applications, their ability to generalize to other scenarios is limited. Consequently, there arises a necessity to train new models for each specific application. This approach results in a multitude of task-specific models for different clinical scenarios, which would incur enormous development costs and lead to a significant waste of resources.

There is a pressing need to build a general-purpose segmentation model for volumetric medical imaging, providing segmentation capabilities applicable to a wide range of anatomical structures and modalities. Currently, Segment Anything Model (SAM) [18], a 2D foundation model built for natural image segmentation, exhibits impressive zero-shot ability to new image distributions and tasks. However, directly applying SAM to 3D volumetric medical images is infeasible, as disclosed in previous evaluations [2, 26]. This infeasibility is not only

due to the domain gap between medical and natural images but also caused by the 2D structure of SAM failing to capture 3D spatial information in volumetric medical images. Here are two straightforward solutions to these issues:

1) Slice-by-Slice Aggregation [6, 20, 25]. This method decomposes each volume into 2D slices, then uses SAM to process each slice, and further aggregates the 2D per-slice results into a 3D prediction. This slice-by-slice manner fails to capture inter-slice correlations, hindering the model's ability to generate consistent 3D predictions. In addition, since each 2D slice necessitates individual prompts for prediction, the large number of 2D slices in a volume yields extensive prompts to obtain the aggregated prediction for such a volume. These prompts, each obtained by per-slice interaction, thereby lead to a burdensome interaction workload.

2) 3D Adapter for Frozen 2D Encoder [3, 10, 37]. In this approach, the 2D encoder layers of SAM are frozen, but 3D adapters are inserted into these frozen 2D layers to enable the model to learn from 3D images. This solution adopts the encoder of SAM to capture 2D information, with adapters only learning about the newly added dimension. This strategy implicitly splits the 3D images along a specific dimension (e.g., two dimensions plus a newly added one), which may curtail their ability to fully model 3D spatial information. Besides, existing adapter-based methods [3, 10, 37] only use small-scale data with limited target types, which cannot provide the tremendous medical knowledge needed for general-purpose segmentation. Table 1 shows the gap between these methods and our work in dataset size and target diversity.

Table 1: Comparison of SAM-based models for volumetric medical images. Our SAM-Med3D employs a fully learnable 3D architecture with large-scale training data, instead of frozen 2D layers with adapters. ❄ and 🔥 denotes frozen and learnable.

Model	Dataset Size	Category	Image Encoder	Prompt Encoder	Mask Decoder
MedLSAM [20]	∼ 25K masks	∼ 50	❄ 2D	❄ 2D	❄ 2D
SAM$^{\text{Med}}$ [31]	-	-	❄ 2D	❄ 2D	❄ 2D
SAM3D [2]	2K masks	14	❄ 2D	-	🔥 **3D**
MA-SAM [3]	≤ 1K masks	≤ 13	❄ 2D + 🔥 Adapter	-	🔥 2D
MSA [37]	12K masks	15	❄ 2D + 🔥 Adapter	🔥 2D	🔥 2D
3DSAM-Adapter [10]	≤ 1K masks	4	❄ 2D + 🔥 Adapter	🔥 **3D**	🔥 **3D**
SAM-Med3D	143K masks	245	🔥 **3D**	🔥 **3D**	🔥 **3D**

Witnessing the success of SAM, we believe that it is crucial to train a general-purpose 3D medical segmentation model using large-scale and diverse 3D medical data. In this paper, we introduce SAM-Med3D, a fully learnable 3D architecture trained on a large-scale volumetric medical dataset using a two-stage procedure. Our SAM-Med3D is a general-purpose promptable segmentation model that needs only a few 3D prompt points to segment diverse anatomical structures and lesions across various modalities, as shown in Fig. 1. We also highlight that SAM-Med3D is a thorough 3D network with all its modules being 3D so that

(a)   (b)

Fig. 2: Overview of SA-Med3D-140K. (a) The word cloud maps for category statistics of all training data. There are 245 categories in our training data. (b) Comparison of counts of images and masks in the 3D medical image datasets. Our dataset consists of 22K 3D images with corresponding 143K 3D masks, while *AMOS* [17], *TotalSegmentator* [36] have less than 2K images, and *BraTS21* [1] has less than 10K masks.

it can better capture varying 3D spatial patterns in the large-scale volumetric medical dataset. Such dataset, termed SA-Med3D-140K, comprises 22K medical images and 143K masks with 245 categories, derived from an amalgamation of 70 public 3D medical image datasets and 24 private datasets. As evident from Fig. 2, it is significantly larger than the largest existing public medical image segmentation datasets like *TotalSegmentator* [36] and *BraTS21* [1]. In addition, we conducted extensive evaluations to verify the promising performance of SAM-Med3D on both seen and unseen tasks (detailed in Sec. 5.1). Specifically, we conducted assessments on our SAM-Med3D and its competitors (SAM and SAM-Med2D) utilizing 16 public datasets. We then thoroughly analyze the results from the perspective of anatomical structures, modalities, and categories to evaluate their performance for general-purpose volumetric medical image segmentation. Lastly, we tested the transferability of the SAM-Med3D on multiple downstream segmentation tasks.

The results reveal three advantages of SAM-Med3D: **1) Better interaction efficiency:** SAM-Med3D shows competitive performance in comparison with state-of-the-art methods but requires fewer prompt points (see Fig. 1). This ensures that, compared to previous methods, physicians can utilize SAM-Med3D more readily and enjoy significantly faster speeds than per-slice interaction. **2) Universal segmentation capabilities:** Compared to previous works, SAM-Med3D exhibits broad segmentation capabilities across a wide array of targets and modalities in volumetric medical images. This versatility underscores the strong adaptability and effectiveness of SAM-Med3D in addressing diverse medical imaging challenges. **3) Promising transfer ability:** When applied to

downstream promptable and semantic segmentation tasks of various new targets, SAM-Med3D exhibits substantial potential as a powerful pre-trained model.

2   Related Works

SAM for Volumetric Medical Images. Just like Language Foundation Models such as GPT-4, Vision Foundation Models (VFM) [18, 28, 30, 35, 41] have demonstrated powerful generalization capabilities. Among them, SAM [18] stands out for promptable image segmentation with zero-shot capabilities in varied visual tasks. Its success in natural images has led to explorations of SAM's potential in volumetric medical imaging, primarily focusing on fine-tuning SAM or incorporating SAM in medical image analysis pipelines. MedLSAM [20] adopts SAM in a two-stage model with a localization model to furnish precise prompt, thereby enhancing SAM's segmentation accuracy. SAM3D [2] leverages a frozen SAM encoder as its image Encoder and processes voxel images slice-by-slice to procure 3D representations, which are subsequently interpreted through a 3D Decoder to generate masks. Besides, other methods try to design 3D adapters to fine-tune SAM for volumetric medical image segmentation. 3DSAM-Adapter [10] and MA-SAM [3] have fashioned 3D adapters tailored for each SAM component, transforming the original SAM structure with 3D convolution to facilitate 3D mask formation. In contrast, MSA [37] retains all the weights of the original SAM and introduces space and depth adapters specifically designed to process 3D spatial information. Nonetheless, these methods encode the vital 3D information based on adapters (only partial parameters are trainable) and only fine-tune SAM on specific limited-scale medical datasets.

Evaluation of SAM in Medical Imaging. Various studies have explored SAM's effectiveness in medical image segmentation. Cheng et al. [5] assess SAM across 12 medical datasets using different prompt types and find its performance generally lagging behind state-of-the-art methods. Huang et al. [14] evaluate SAM's zero-shot capability on 52 datasets with three prompt types and observe the subpar performance of SAM. Deng et al. [7] focus on tumor and tissue segmentation, noting SAM's better performance with larger connected entities, while Zhou et al. [40] identify potential improvements in colonoscopy polyp segmentation using SAM without prompts. Hu et al. [13] investigate SAM in multi-phase liver tumor segmentation in CT scans, finding increased effectiveness with more point prompts. These studies highlight the importance of understanding SAM's capabilities in medical imaging, offering insights to refine SAM for medical segmentation tasks. Complementing these efforts, our research provides a comprehensive evaluation of several SAM-based methods, including our SAM-Med3D, on volumetric medical images.

3   SA-Med3D-140K Dataset

3.1   Data Acquisition

We constructed a large-scale 3D volumetric medical image segmentation dataset, SA-Med3D-140K, based on 70 public and 24 private datasets. SA-Med3D-140K contains 22K medical images and 143K masks which is probably the largest volumetric medical image segmentation dataset to date. The dataset covers 28 modalities (i.e. CT, US, and 26 MR sequences) with 6 major anatomical structures. As shown in Fig. 2, there are over 240 categories of targets including both organs and lesions in the training set. It is about 20 times larger than the largest well-known medical image segmentation dataset for organ, *TotalSegmentator* [36], at the image level. Details of dataset are in Supplementary.

3.2   Data Pre-processing

To standardize the diverse data from multiple datasets and control the data quality, we clean and process all the collected data in the following four steps:
1) Data cleaning based on target shapes: We first summarize the sizes of all targets in each medical image. Then, we removed all masks with a physical size below 1 cm^3 or with any single dimension shorter than 1.5 cm to ensure the visibility of target masks under the target spacing 1.5 mm.
2) Data cleaning based on volume sizes: We convert multi-class masks into one-hot formats, calculate the total volume sizes of the foreground, and discard masks where the background exceeds 99% of the volume.
3) Noise reduction based on connected domains: For the remaining masks, our pipeline enhances label quality by removing domains smaller than the top-5 largest connected domains.
4) Ambiguity reduction based on symmetry: We directly manipulate the data by segregating symmetric target masks, such as dividing *Kidney* into *Left Kidney* and *Right Kidney*. This approach contrasts with SAM's method of generating multiple predictions per prompt to address ambiguity; we choose direct data manipulation because we observe that medical image masks typically present less inherent ambiguity.

3.3   Data Splitting

To train and evaluate the performance of our model, we partitioned the SA-Med3D-140K into a training set of 131K masks and a validation set of 12K masks. The training set incorporates 58 public datasets along with all private data. Conversely, the validation set is composed of 16 public benchmark datasets, some of which have their official training sets included within our training set.

This splitting strategy is mainly designed to assess the model's ability to generalize across diverse data source and modalities. Among the 16 datasets in the validation set, all annotated images from 12 datasets have been used for evaluation, implying that 58% of the cases in our validation set have data sources

that are unseen in the training set. Besides, we divided all ultrasound (US) data into the validation set to assess the cross-modality generalization ability. Details of data splitting are in Supplementary.

4 Method

4.1 Devise SAM-Med3D

Table 2: Preliminary experiment comparing three modifications of SAM for volumetric medical images. All models are trained on the training set of *AMOS* and evaluated on seen and unseen targets from the validation/test set of *AMOS* and *Totalsegmentator*. The Dice scores are averaged at prompt points number of 1 and 10 for brevity.

Method	Dice (%)		
	Seen	Unseen	Overall
3D adapter w/ frozen SAM [38]	49.08	38.60	43.84
Fine-tuning w/ SAM pre-training	69.13	18.43	43.78
Training from scratch on 3D data	55.79	34.63	45.21

While SAM excels in prompt-based segmentation on 2D natural images, the 2D architecture of SAM fails to capture 3D spatial information of volumetric images, leading to sub-optimal performance. Modifying SAM for 3D volumetric medical images remains an open question, and there are three potential methods to address this issue:

1) 3D Adapter with Frozen SAM: Similar to image-to-video adaptation [38], this approach involves training additional modules (i.e. adapter) on top of the frozen 2D architecture of SAM to learn 3D medical knowledge.

2) Fine-tuning with SAM Pre-training: This method repurposes the 2D pre-trained weights for a 3D architecture (i.e. replicating 2D weights of convolution to formulate 3D weights), and conducts full fine-tuning on 3D medical images to acquire new knowledge.

3) Training from Scratch on 3D Data: This approach entails designing an entirely 3D structure and training from scratch on 3D volumetric medical data, without utilizing SAM's pre-trained weights.

Each method has its pros and cons. The first method benefits from the pre-trained weights of SAM for a speedy convergence, but 3D adapters, built upon the frozen 2D structure of SAM, may be misled by the prior knowledge that is highly biased to 2D natural images. Moreover, most of the additional adapters only learn knowledge for the new z-axis, which is sub-optimal for volumetric medical images with three axes. The second one can also converge fast due to the use of pre-trained weights from SAM. Furthermore, it can directly capture 3D information for its 3D architecture. However, the 2D-to-3D weight transition in this method might further break down the prior knowledge of SAM, which is harmful to the generalization ability the original SAM brings. Such corrupt
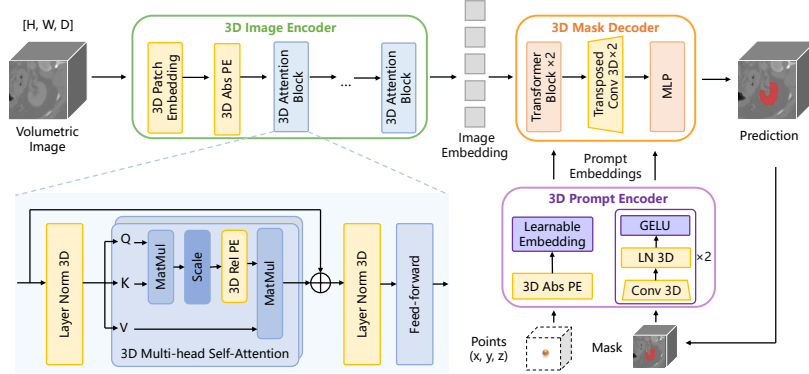
Fig. 3: The fully 3D architecture of our SAM-Med3D, encompassing a 3D image encoder, 3D prompt encoder, and 3D mask decoder. 3D positional encoding (PE) and 3D layers like convolution and layer normalization are employed to construct it.

knowledge may further lead to the lack of strong generalization ability of the fine-tuned model. The third method is advantageous as it is free from biased knowledge from 2D natural images, making it more adaptable for 3D tasks. Yet, it converges much slower and increases the training difficulty.

To identify the optimal solution, we compare them via preliminary experiments. We train a model for each approach on the training set of *AMOS* and evaluate them on both seen and unseen targets, the former from the validation set of *AMOS* and the latter from the test set of *Totalsegmentator*. We average the Dice scores of two prompt settings (1 and 10 prompt points) for brevity.

As illustrated in Table 2, the 3D adapter shows suboptimal performance on seen targets, while full fine-tuning struggles to generalize on unseen targets; training from scratch emerges as a better trade-off, exhibiting superior average performance. Based on the results, we opted for training a fully 3D architecture from scratch on 3D medical data. The schematics of our SAM-Med3D architecture are in Fig. 3. Our SAM-Med3D leverages 3D positional encoding and 3D layers (such as convolution and layer normalization) to directly process 3D images and 3D prompts. This pure 3D design can better capture 3D spatial information, and significantly reduce required prompt point numbers. More details of our SAM-Med3D are in Supplementary.

4.2 Train SAM-Med3D in Two Stages

Inspired by the prevalence of the pre-training and fine-tuning paradigm in Large Language Models (LLM), we adopted a two-stage procedure in SAM-Med3D model training. We kept most training setting the same for both stages (except the iterations), with the major difference being the training data.

Stage 1: Pre-training. At the first stage, we utilize all 131K data in the training set of SA-Med3D-140K to construct a powerful pre-trained model. We train the model for 800 epochs until it essentially converges. At this point, SAM-

Med3D exhibits performance on common targets that is on par with SAM-Med2D, yet it still falls short on challenging targets.

Stage 2: Fine-tuning. At this stage, we filter about 75K high-quality masks from the entire training dataset for the 2nd-stage fine-tuning. The impact of different fine-tuning datasets is discussed in Sec. 5.1. More details of training are in Supplementary.

4.3   Evaluate General-Purpose Models

Based on the elaborated validation set of the general-purpose dataset SA-Med3D-140K, we undertake a comprehensive evaluation of SAM, SAM-Med2D (the SOTA finetuned SAM for 2D medical images), and our SAM-Med3D, aiming to set a benchmark for promptable segmentation tasks on 3D medical images. Notably, unlike task-specific methods that train multiple models for various tasks, all the compared methods employ a single model for all different tasks. These models need prompts to yield predictions. In terms of the prompt, we simulated an interactive segmentation clinical scenario using point prompt mode across all models, with the first prompt randomly sampled from the foreground and subsequent points from the error region.

In our evaluation, SAM-Med2D and SAM resize each slice to target resoltions during the per-slice interaction, while SAM-Med3D uses a patch-based inference for 3D volumes. For volumes over 128, SAM-Med3D starts by cropping a 128-sized patch around the initial point. If the patch's edge has prediction, the model extends its inference in that direction with a 50% overlapping sliding window, with the prompt still visible at the edge. This approach balances SAM-Med3D's speed and accuracy. Notably, for most targets in our validation set, SAM-Med3D needs to predict only one or two patches.

5   Experiments

5.1   Quantitative Evaluation

Overall Performance Table 3 presents the performance of SAM, SAM-Med2D, and SAM-Med3D on our evaluation dataset introduced in Sec. 3.3. The first two rows in the table reveal that SAM fine-tuned with medical domain knowledge (i.e. SAM-Med2D) clearly outperforms SAM by 22.12% and 10.82% for seen and unseen targets. In contrast, our SAM-Med3D (the 3rd row) shows an even more pronounced performance leap, registering an improvement of 60.12% in overall Dice scores. Crucially, SAM-Med3D consistently exceeds both SAM and SAM-Med2D across varied prompt point counts, although it needs much fewer points.

Another significant advantage of our SAM-Med3D is its efficiency. Since the fully 3D architecture of SAM-Med3D enables greater throughput and reduces interaction burden, it operates at only 1% to 26% of the inference time required by SAM under different sizes of targets. Even when excluding human

Table 3: Comparison of different methods on our validation set (including 16 datasets, detailed in Sec. 3.1). Here, N denotes the count of slices containing the target ($10 \leq N \leq 200$). τ represents the interaction time (usually more than 1 second).

Model	Prompt	Inference Time (s)	Dice (%)		
			Seen	Unseen	Overall
SAM	N pts	$N(\tau + 0.13)$	16.79	11.73	16.15
SAM-Med2D	N pts	$N(\tau + 0.04)$	38.91	22.55	36.83
SAM-Med3D	1 pt	$\tau + 2$	81.98	37.02	76.27
SAM	$3N$ pts	$3N(\tau + 0.19)$	34.61	15.94	32.24
SAM-Med2D	$3N$ pts	$3N(\tau + 0.07)$	51.46	29.70	48.70
SAM-Med3D	3 pts	$3\tau + 3$	84.14	43.80	79.02
SAM	$5N$ pts	$5N(\tau + 0.25)$	49.39	21.86	45.89
SAM-Med2D	$5N$ pts	$5N(\tau + 0.10)$	51.89	30.41	49.17
SAM-Med3D	5 pts	$5\tau + 4$	84.62	46.26	79.75
SAM-Med3D	10 pts	$10\tau + 6$	85.19	49.92	80.71

interaction time (i.e. $\tau = 0$), SAM costs more inference time when N surpasses 20 (N denotes the number of slices containing the target). This indicates that for objects with diameters greater than 3 cm (assuming a spacing of 1.5 mm), SAM-Med3D emerges as a more efficient choice. This enhancement in efficiency makes it particularly suitable for real-world medical applications where processing larger volumes quickly is crucial.

Table 4: Comparison of task-specific and general-purpose methods on 6 public benchmarks (task-specific models are trained on the training set of each dataset). N denotes the count of slices containing the target and $10 \leq N$. Items marked with * represent the unseen data source or modality in the training set of SA-Med3D-140K.

Dataset	Modality	Task-specific		General-purpose			
		UNETR [11]	nnU-Net [16]	SAM-Med2D [6] (N pts)	SegVol [8] (pt+text)	Ours (1 pt)	Ours (10 pts)
Totalsegmentator [36]	CT	75.05	85.22	38.26	-	84.68	**87.59**
KiTS21 [12]	CT	70.75	75.32	68.74	-	72.06	**75.37**
AMOS-CT [17]	CT	78.33	**88.87**	49.61	-	79.94	83.99
AMOS-MR [17]	MR	76.29	**86.92**	45.53	-	75.41	81.13
BTCV* [19]	CT	78.99	81.92	50.05	73.81	79.17	**83.01**
TDSC-ABUS23* [33]	US*	-	45.08	49.39	-	36.08	**54.35**

Comparison with Task-specific Models We conducted a comparison between task-specific models and general-purpose model to show the gap between specialists and generalist on public benchmarks. For task-specific models, we select the representative CNN-based model nnU-Net [16] and Transformer-based model UNETR [11] to conduct experiments. We trained task-specific models on the training set of each benchmark respectively in our evaluations. For general-purpose models, we tested them using the official weights provided.

Table 4 summarizes the performance of 5 models on the validation/test set of 6 public benchmarks. We found that SAM-Med3D consistently achieved competitive Dice scores across all datasets compared to task-specific models. When provided with 10 points ($10 \leq N$), SAM-Med3D outperformed all other competitors on 4 benchmarks, including unseen data sources and modalities. We also noted that on the unseen modality Ultrasound (US), the increase of prompt point numbers can bring more significant performance improvements for SAM-Med3D. The performance of SAM-Med2D also surpasses that of the SOTA 3D task-specific models because its training data includes 2D US slices.

Table 5: Comparison from the perspective of anatomical structure and lesion on our validation set (including 16 datasets, detailed in Sec. 3.1). Abd&Tho represents Abdominal and Thorax targets. N denotes the count of slices containing the target ($10 \leq N \leq 200$).

Model	Prompt	Seen						Unseen	
		Abd&Tho	Bone	Brain	Cardiac	Muscle	Lesion	Organ	Lesion
SAM	N pts	19.93	17.85	29.73	8.44	3.93	11.56	12.14	8.88
SAM-Med2D	N pts	50.47	32.70	36.00	40.18	43.85	24.90	19.36	44.87
SAM-Med3D	1 pt	80.76	83.38	43.74	87.12	89.74	58.06	35.99	44.22
SAM	$3N$ pts	31.97	42.07	40.42	18.94	8.46	24.52	16.18	14.21
SAM-Med2D	$3N$ pts	58.94	47.48	54.57	47.06	56.05	42.60	27.25	46.87
SAM-Med3D	3 pts	84.06	84.74	53.34	89.71	91.01	61.73	42.20	55.02
SAM	$5N$ pts	43.45	59.87	49.85	33.57	16.94	38.86	21.67	23.22
SAM-Med2D	$5N$ pts	59.52	47.81	55.02	47.38	56.36	43.90	27.95	47.66
SAM-Med3D	5 pts	84.68	85.02	56.86	90.36	91.29	62.94	44.51	58.46
SAM-Med3D	10 pts	85.42	85.34	61.27	90.97	91.62	64.80	48.10	62.72

Evaluation on Different Anatomical Structures Table 5 summarizes the results of SAM, SAM-Med2D, and our SAM-Med3D across different anatomical structures and lesions. With only N points (1 point/slice), SAM frequently exhibits poor performance. This issue is not readily resolved even with additional points ($3N$ or $5N$ points), especially in anatomical structures with unclear boundaries such as cardiac and muscle regions. In contrast, SAM-Med2D and our SAM-Med3D, infused with medical knowledge, are more effective in accurately identifying targets with just a single point per slice or volume. Notably, with only one point, SAM-Med3D achieves significant better performance than SAM-Med2D across a range of seen and unseen targets, except the unseen lesion. This outlier occurs because one of the unseen lesion modality (i.e. Ultrasound) for SAM-Med3D, is already familiar to SAM-Med2D. As the number of prompt points is increased, our SAM-Med3D distinctly leads for all examined anatomical structures and lesions, showcasing its robustness and versatility.

In our evaluation of seen targets across various anatomical structures, we found that segmenting the brain structures and lesions was more challenging
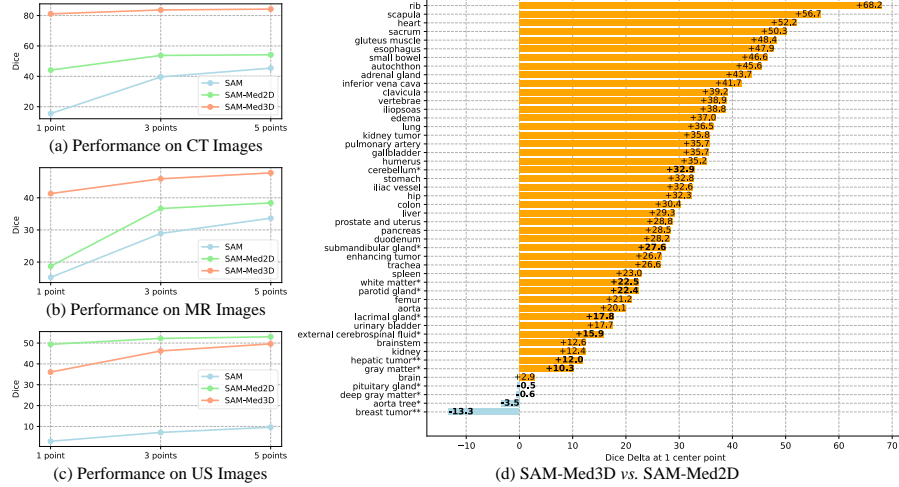
Fig. 4: (a-c) Comparison across different modalities with varying numbers of points. Despite not being trained on the US modality like SAM-Med2D, SAM-Med3D still shows competitive performance. (d) Comparison of the Dice score between SAM-Med3D and the 2D fine-tuned SAM, SAM-Med2D [6] across 44 major organs and 5 kinds of lesions. ∗ and ∗∗ represent unseen organs and lesions.

for all models. We noted that for the brain structures, with sufficient prompts, SAM-Med2D begins to perform competitively with SAM-Med3D; however, for lesions, additional prompts did not significantly assist SAM-Med2D in reducing the gap with SAM-Med3D. We attribute this difference to lesions being more dependent on prompts. This idea is corroborated in the case of unseen targets. In comparison to unseen organs, the increase in prompt points has a significantly larger effect on SAM-Med3D's recognition of unseen lesions.

Evaluation on Different Modalities We compared the three methods for three dominant modalities in volumetric medical imaging (CT, MR, and US) across various numbers of prompt points. For brevity, we averaged the results from all MR sequences to present them effectively as MR results. As shown in Fig. 4, SAM underperforms in segmenting all three volumetric modalities with just one point per slice, yielding Dice scores below 20%. Although SAM shows better results on all three modalities with more prompt points for each slice, its performance is still far away from the other two methods.

It is observed that SAM-Med3D performs better with two mainstream imaging modalities (CT and MR). Even in the case of the unseen modality US, SAM-Med3D exhibits comparable performance when provided with sufficient prompts.

Evaluation on Major Organs and Lesions We have organized all the categories and classified them according to the specific organ or lesion that they belong to. For example, the results for *Left Kidney* and *Right Kidney* are averaged into *Kidney*. By this means, we have identified 44 major organs and 5 kinds of lesions. Fig. 4 (d) shows that SAM-Med3D using 1 point outperforms SAM-Med2D with N points in 45 targets out of 49, achieving up to +68.2% improvement. Consistent with the observation of anatomical structures in Sec. 5.1, SAM-Med3D shows greater advantages in segmenting bones, cardiac structures and muscles, maybe benefiting from their sensitivity to the 3D spatial correlations. For brain structures and unseen targets, SAM-Med3D exhibits a relatively smaller advantage.

Table 6: Transferability evaluation of semantic segmentation. We trained the SOTA ViT-based model (i.e. UNETR [11]), with and without our SAM-Med3D pre-trained ViT encoder, to assess the benefits of pre-training.

Pre-train	AMOS [17]	Totalsegmentator [36]	CAS2023 [4]	SEG.A.2023 [29]	Avg.
-	76.29	82.67	86.34	87.05	83.09
SAM-Med3D	81.92	85.17	88.39	87.57	85.79

Table 7: Transferability evaluation of promptable segmentation. The first row displays baseline results from SAM-Med3D only with the 1st-stage pre-training. The impact of various fine-tuning datasets during the 2nd-stage fine-tuning is assessed on our validation set (excluding *FeTA22* [27] for fairness). The last row denotes SAM-Med3D fine-tuned with high-quality data selected from 44 datasets in SA-Med3D-140K.

Fine-tune Target	Fine-tune Dataset	Seen							Unseen	
		Abd&Tho	Bone (vertebrae)	Bone (other)	Brain	Cardiac	Muscle	Lesion	Organ	Lesion
-	-	58.61	66.05	28.08	56.34	68.50	69.45	47.87	27.80	48.44
organ	*AMOS* [17]	67.23	58.59	30.51	59.72	71.89	77.39	50.47	38.30	37.43
vertebrae	*Verse20* [23]	35.08	78.84	23.77	7.77	61.02	37.75	35.42	18.38	24.69
brain	*FeTA22* [27]	50.28	51.14	25.34	62.56	62.95	43.70	45.89	44.81	24.73
all	44 datasets	85.42	85.98	84.90	61.91	90.97	91.62	64.80	48.10	62.72

Evaluation on Transferability We test the transferability of SAM-Med3D as a pre-train model on two important downstream tasks:

For semantic segmentation, we choose two frequently-used benchmarks (i.e. *AMOS* [17], *Totalsegmentator* [36]) and two unseen datasets from the MICCAI 2023 Challenge (i.e. *CAS2023* [4], *SEG.A.2023* [29]). The image encoder of SAM-Med3D is used as the pre-trained feature extractor for UNETR [11], a leading-edge ViT-based model in medical segmentation tasks. As depicted in Table 6, when fine-tuned with pre-trained ViT encoder from SAM-Med3D, UNETR demonstrates a substantial performance boost compared to these without pre-training, achieving a maximum improvement of 5.63% in the Dice score.
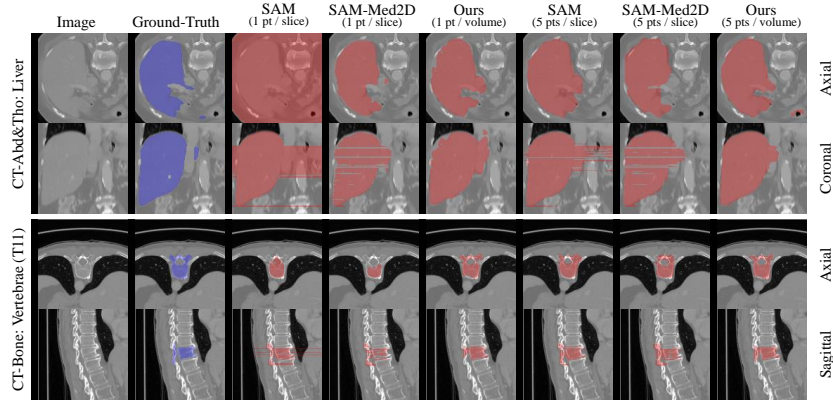
Fig. 5: Visualization of SAM, SAM-Med2D, and our SAM-Med3D across diverse anatomical structures and modalities for 1 or 5 points. We present both axial and coronal/sagittal views to illustrate the 3D results comprehensively.

Consequently, SAM-Med3D demonstrates the substantial potential to establish itself as a powerful pre-trained ViT encoder for multiple downstream tasks.

For promptable segmentation, we carried out experiments as outlined in Section 4.2. We assessed the impact of various fine-tuning datasets during the second-stage fine-tuning from the viewpoint of anatomical structures and lesions. As demonstrated in Table 7, fine-tuning on specific datasets significantly enhances the model's expertise in targeted entities; however, this specialization inadvertently biases the model, leading to decreased performance on non-target entities. Conversely, fine-tuning on carefully selected, high-quality data further strengthens the model's general-purpose segmentation performance.

5.2   Qualitative Evaluation

To qualitatively compare the performance of SAM-Med3D and other methods, we visualize the predicted masks across different point numbers in Fig. 1 and 5. Based on these visualization results, we highlight two key observations: 1) SAM-Med3D requires fewer prompts. 2) SAM-Med3D shows better inter-slice consistency than all other methods.

6   Conclusion

In this paper, we introduce a general-purpose promptable model for volumetric medical image segmentation, dubbed SAM-Med3D. SAM-Med3D uses only a few prompt points to correctly segment both seen and unseen anatomical structures and modalities, underpinning its general-purpose segmentation ability. Such an impressive ability is obtained by training this fully learnable 3D network on a large-scale 3D medical dataset in two stages. Extensive evaluations on 16 public datasets verify the efficacy and efficiency of our SAM-Med3D in various clinical

settings, e.g., a 60.12% improvement over SAM with one point per volume. Our model also shows promising potential to serve as a powerful pre-trained transformer model for multiple downstream segmentation tasks.

References

1. Baid, U., Ghodasara, S., Bilello, M., Mohan, S., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., Prevedello, L.M., Rudie, J.D., Sako, C., Shinohara, R.T., Bergquist, T., Chai, R., Eddy, J.A., Elliott, J., Reade, W., Schaffter, T., Yu, T., Zheng, J., Annotators, B., Davatzikos, C., Mongan, J., Hess, C., Cha, S., Villanueva-Meyer, J.E., Freymann, J.B., Kirby, J.S., Wiestler, B., Crivellaro, P., Colen, R.R., Kotrotsou, A., Marcus, D.S., Milchenko, M., Nazeri, A., Fathallah-Shaykh, H.M., Wiest, R., Jakab, A., Weber, M., Mahajan, A., Menze, B.H., Flanders, A.E., Bakas, S.: The RSNA-ASNR-MICCAI brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. CoRR **abs/2107.02314** (2021), https://arxiv.org/abs/2107.02314
2. Bui, N.T., Hoang, D.H., Tran, M.T., Le, N.: Sam3d: Segment anything model in volumetric medical images. arXiv preprint arXiv:2309.03493 (2023)
3. Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al.: Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. arXiv preprint arXiv:2309.08842 (2023)
4. Chen, H., Zhao, X., Sun, H., Dou, J., Du, C., Yang, R., Lin, X., Yu, S., Liu, J., Yuan, C., Balu, N.: Cerebral artery segmentation challenge. In: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) (2023)
5. Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K.: Sam on medical images: A comprehensive study on three prompt modes (2023)
6. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
7. Deng, R., Cui, C., Liu, Q., Yao, T., Remedios, L.W., Bao, S., Landman, B.A., Wheless, L.E., Coburn, L.A., Wilson, K.T., Wang, Y., Zhao, S., Fogo, A.B., Yang, H., Tang, Y., Huo, Y.: Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging (2023)
8. Du, Y., Bai, F., Huang, T., Zhao, B.: Segvol: Universal and interactive volumetric medical image segmentation. arXiv preprint arXiv:2311.13385 (2023)
9. Gatidis, S., Früh, M., Fabritius, M., Gu, S., Nikolaou, K., La Fougère, C., Ye, J., He, J., Peng, Y., Bi, L., et al.: The autopet challenge: Towards fully automated lesion segmentation in oncologic pet/ct imaging (2023)
10. Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.A., Dou, Q.: 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. arXiv preprint arXiv:2306.13465 (2023)
11. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
12. Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., Shoshan, Y., Gilboa-Solomon, F., George, Y., Yang, X., Zhang, J., Zhang, J., Xia, Y., Wu, M., Liu, Z., Walczak, E., McSweeney, S., Vasdev, R., Hornung, C., Solaiman, R., Schoephoerster, J., Abernathy, B.,

Wu, D., Abdulkadir, S., Byun, B., Spriggs, J., Struyk, G., Austin, A., Simpson, B., Hagstrom, M., Virnig, S., French, J., Venkatesh, N., Chan, S., Moore, K., Jacobsen, A., Austin, S., Austin, M., Regmi, S., Papanikolopoulos, N., Weight, C.: The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct (2023)

13. Hu, C., Xia, T., Ju, S., Li, X.: When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation (2023)

14. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., Chi, H., Hu, X., Fan, D.P., Dong, F., Ni, D.: Segment anything model for medical images? (2023)

15. Huang, Z., Wang, H., Deng, Z., Ye, J., Su, Y., Sun, H., He, J., Gu, Y., Gu, L., Zhang, S., et al.: Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. arXiv preprint arXiv:2304.06716 (2023)

16. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

17. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)

18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)

19. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge. vol. 5, p. 12 (2015)

20. Lei, W., Wei, X., Zhang, X., Li, K., Zhang, S.: Medlsam: Localize and segment anything model for 3d medical images. arXiv preprint arXiv:2306.14752 (2023)

21. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)

22. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. arXiv preprint arXiv:2301.00785 (2023)

23. Löffler, M.T., Sekuboyina, A., Jacob, A., Grau, A.L., Scharr, A., El Husseini, M., Kallweit, M., Zimmer, C., Baum, T., Kirschke, J.S.: A vertebral segmentation dataset with fracture grading. Radiology: Artificial Intelligence **2**(4), e190138 (2020)

24. Luo, X., Wang, G., Song, T., Zhang, J., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S.: Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. Medical image analysis **72**, 102102 (2021)

25. Ma, J., Wang, B.: Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2023)

26. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. Medical Image Analysis **89**, 102918 (2023)

27. Payette, K., de Dumast, P., Kebiri, H., Ezhov, I., Paetzold, J.C., Shit, S., Iqbal, A., Khan, R., Kottke, R., Grehten, P., et al.: An automatic multi-tissue human fetal

brain segmentation benchmark using the fetal tissue annotation dataset. Scientific data **8**(1), 167 (2021)

28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
29. Radl, L., Jin, Y., Pepe, A., Li, J., Gsaxner, C., Zhao, F.h., Egger, J.: Avt: Multicenter aortic vessel tree cta dataset collection with ground truth segmentation masks. Data in brief **40**, 107801 (2022)
30. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation (2021)
31. Wang, C., Li, D., Wang, S., Zhang, C., Wang, Y., Liu, Y., Yang, G.: Sammed: A medical image annotation framework based on large vision model. arXiv preprint arXiv:2307.05617 (2023)
32. Wang, H., Yi, H., Liu, J., Gu, L.: Integrated treatment planning in percutaneous microwave ablation of lung tumors. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 4974–4977. IEEE (2022)
33. Wang, K.: Tumor detection, segmentation and classification challenge on automated 3d breast ultrasound (abus) 2023 (2023), `https://tdsc-abus2023.grand-challenge.org/TDSC-ABUS2023/`
34. Wang, X., Jiang, L., Li, L., Xu, M., Deng, X., Dai, L., Xu, X., Li, T., Guo, Y., Wang, Z., Dragotti, P.L.: Joint learning of 3d lesion segmentation and classification for explainable covid-19 diagnosis. IEEE Transactions on Medical Imaging **40**(9), 2463–2476 (2021). `https://doi.org/10.1109/TMI.2021.3079709`
35. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Segmenting everything in context. arXiv preprint arXiv:2304.03284 (2023)
36. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
37. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
38. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: Aim: Adapting image models for efficient video action recognition. arXiv preprint arXiv:2302.03024 (2023)
39. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1195–1204 (2021)
40. Zhou, T., Zhang, Y., Zhou, Y., Wu, Y., Gong, C.: Can sam segment polyps? (2023)
41. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once (2023)