Medical Image Analysis 102 (2025) 103547

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Medical SAM adapter: Adapting segment anything model for medical image segmentation

Junde Wu [a], Ziyue Wang [b], Mingxuan Hong [a], Wei Ji [c], Huazhu Fu [d], Yanwu Xu [e], Min Xu [f,g], Yueming Jin [a,b,*]

[a] Department of Biomedical Engineering, National University of Singapore, Singapore
[b] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[c] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada
[d] Institute of High-Performance Computing, Agency for Science, Technology and Research, 138632, Singapore
[e] Singapore Eye Research Institute, Singapore
[f] Computer Vision Department, Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[g] Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA 15213, United States of America

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Medical Image Segmentation
Efficient Learning
Fine-Tuning

ABSTRACT

The Segment Anything Model (SAM) has recently gained popularity in the field of image segmentation due to its impressive capabilities in various segmentation tasks and its prompt-based interface. However, recent studies and individual experiments have shown that SAM underperforms in medical image segmentation due to the lack of medical-specific knowledge. This raises the question of how to enhance SAM's segmentation capability for medical images. We propose the Medical SAM Adapter (Med-SA), which is one of the first methods to integrate SAM into medical image segmentation. Med-SA uses a light yet effective adaptation technique instead of fine-tuning the SAM model, incorporating domain-specific medical knowledge into the segmentation model. We also propose Space-Depth Transpose (SD-Trans) to adapt 2D SAM to 3D medical images and Hyper-Prompting Adapter (HyP-Adpt) to achieve prompt-conditioned adaptation. Comprehensive evaluation experiments on 17 medical image segmentation tasks across various modalities demonstrate the superior performance of Med-SA while updating only 2% of the SAM parameters (13M). Our code is released at https://github.com/KidsWithTokens/Medical-SAM-Adapter.

1. Introduction

Recently, the Segmentation Anything Model (SAM) (Kirillov et al., 2023) has gained significant attention as a powerful and versatile vision segmentation model. It can generate diverse and detailed segmentation masks based on user prompts. Despite its strong performance over natural images, recent studies also show (Deng et al., 2023b; Roy et al., 2023) that it reaches subpar performance on medical image segmentation. Making medical image segmentation interactive, such as employing techniques like SAM, holds immense clinical value. An interactive system can prioritize areas of interest as indicated by the clinicians, providing them with a more immersive and personalized experience. For instance, in a single fundus image, there are often overlapping and intricately intertwined structures such as vessels, optic discs, optic cups, and macula. Interactive segmentation can greatly assist clinicians in efficiently distinguishing target tissues from these complex structures. Meanwhile, considering the difficulty in acquiring large-scale annotated medical datasets, it becomes crucial to adopt available foundational models like SAM for clinical utilization.

SAM's limited performance on medical images is due to its lack of medical-specific knowledge, including challenges like low image contrast, ambiguous tissue boundaries, and tiny lesion regions. The state-of-the-art (SOTA) approach to address this issue is fully fine-tuning the vanilla SAM model specifically on medical data (Ma and Wang, 2023), which is quite costly in terms of both computation and memory footprint. Additionally, it is doubtful whether fully fine-tuning is necessary, as previous studies have shown pre-trained visual models have strong transferability to medical images (Raghu et al., 2019; Xie and Richmond, 2018).

This paper attempts to adapt the well-trained SAM to medical image segmentation with minimum effort. Our primary goal is to develop a simple yet effective adapter technology that can efficiently adapt SAM to the medical field while ensuring strong generalization. Technically,

* Corresponding author.
E-mail address: ymjin@nus.edu.sg (Y. Jin).

https://doi.org/10.1016/j.media.2025.103547
Received 14 August 2024; Received in revised form 31 December 2024; Accepted 7 March 2025
Available online 19 March 2025

we choose to fine-tune the pre-trained SAM using a parameter-efficient fine-tuning (PEFT) technique called Adaption (Hu et al., 2021). Adaption has been a popular and widely used technology in natural language processing (NLP) to fine-tune the fundamental pre-trained model for various downstream tasks. The main idea of Adaption is to insert Adapter modules with partial parameters into the original model and only update a small number of additional Adapter parameters while keeping the large pre-trained model frozen. However, directly applying the Adaption technique to the medical scenario is not that straightforward. The first challenge arises from the image modality. Unlike natural images, many medical images are 3D, such as CT and MRI scans. It is unclear how to adapt the 2D SAM model for 3D medical image segmentation. Secondly, while Adaption has been successful in NLP, there is limited research on applying it to visual models, especially interactive visual models like SAM. In interactive visual models, user-provided visual prompts play a crucial role in the final prediction. How to incorporate Adaption with these important visual prompts remains unexplored.

To overcome these challenges, we propose a novel adaptation framework called Medical SAM Adapter (Med-SA). A Space-Depth Transpose (SD-Trans) technique is introduced in Med-SA to achieve 2D to 3D adaptation. In SD-Trans, we transpose the spatial dimension of input embedding to the depth dimension, allowing the same self-attention blocks to process different dimensional information given various inputs. Then we propose a Hyper-Prompting Adapter (HyP-Adpt) to enable prompt-conditioned adaptation, in which we use the visual prompt to generate a series of weights that can be applied to the adaptation embedding efficiently, facilitating wide and deep prompt-adaptation interactions.

We conduct comprehensive evaluation experiments covering 17 medical image segmentation tasks across various modalities from 5 datasets, including CT, MRI, ultrasound images, fundus images, and dermoscopic images (We refer the "task" as segmenting an specific organ/tissue in our interactive segmentation setting.). By updating merely 2% extra parameters of the total SAM parameters, Med-SA outperforms both SAM and fully fine-tuned SAM (MedSAM) (Ma and Wang, 2023). It also surpasses several SOTA methods that are tailor-designed for medical image segmentation, such as nnUNet, TransUNet, UNetr, and Swin-UNetr. Our work is ***one of the first explorations of SAM in the medical domain***, which has received great impact and high interest in the medical image analysis field. It has been widely recognized as an important architecture that facilitates the development of foundation models for medical image segmentation (Huang et al., 2024; Zhang and Metaxas, 2023; Qiu et al., 2023; Zhang et al., 2024).

- We present the Adaption approach for general medical image segmentation. Our framework, Med-SA, is a simple yet powerful extension of the SAM architecture, substantially enhancing its capabilities for medical applications while updating a mere 2% of the total parameters.
- We propose SD-Trans to enable the segmentation of high-dimensional medical data, addressing the challenge posed by medical image modalities.
- We propose HyP-Adpt to facilitate prompt-conditioned Adaption, acknowledging the importance of user-provided prompts in the medical domain.
- We have conducted extensive and comprehensive comparisons with numerous advanced methods, which we diligently gathered from four aspects. Our extensive experiments on 17 medical image segmentation tasks with various image modalities establish Med-SA's superiority over SAM and previous SOTA methods. On the widely-used abdominal multi-organ segmentation BTCV benchmark, Med-SA outperforms Swin-UNetr by 2.9%, vanilla SAM by 34.8%, and fully fine-tuned SAM (MedSAM) by 9.4%.

2. Related work

2.1. Medical image segmentation

Deep learning methods have been extensively explored for medical image segmentation. In CNN-based approaches, U-Net (Ronneberger et al., 2015) introduced the innovative skip connection structure, inspiring a series of works that achieved impressive performance (Çiçek et al., 2016; Zhou et al., 2018; Isensee et al., 2021; Chu et al., 2021). Following the success of Transformers in computer vision tasks (Dosovitskiy, 2020), transformer-based and transformer-CNN hybrid models have also been designed in the medical domain (Wang et al., 2021b; Hatamizadeh et al., 2022a; Chen et al., 2024; Xie et al., 2021), such as TransUNet (Chen et al., 2021a) and UNetr (Hatamizadeh et al., 2022b). Recently, the emergence of new backbones, such as diffusion models (Ho et al., 2020; Song et al., 2020) and Mamba (Gu and Dao, 2023), has attracted attention, leading to the development of segmentation methods based on these architectures (Wolleb et al., 2021; Wu et al., 2022b; Kim et al., 2022; Xing et al., 2023; Liu et al., 2024) like Segdiff (Amit et al., 2021) and U-mamba (Ma et al., 2024). However, these methods are limited to the semantic categories predefined in the training set and lack the ability to interact with humans, thus cannot flexibly segment organs or tissues based on user needs.

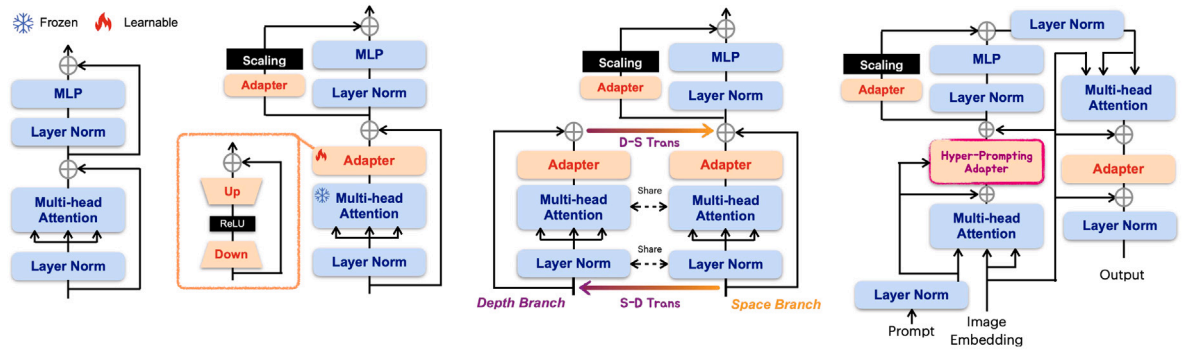
2.2. Interactive segmentation

Interactive segmentation has a rich history, initially regarded as an optimization technique by researchers (Grady, 2006; Gulshan et al., 2010). The pioneering work of DIOS (Xu et al., 2016) revolutionized interactive segmentation by integrating deep learning and incorporating positive and negative clicks as distance maps. Subsequent studies (Li et al., 2018) focused on addressing uncertainty by predicting multiple potential results and enabling either a selection network or the user to choose among them. CDNet (Chen et al., 2021b) further enhanced interactive segmentation by incorporating self-attention to generate more consistent predictions. Recently, SAM (Roy et al., 2023) demonstrated the significant impact of interactive segmentation. Several studies have adopted the SAM framework for medical image segmentation tasks (Ma and Wang, 2023; Zhang et al., 2023b; Cheng et al., 2023; Deng et al., 2023a; Zhang et al., 2024). However, the majority require extensive tuning of the SAM parameters, leading to significant increases in training time and GPU memory usage. These demands are challenging to meet in many computational environments.

2.3. Parameter-efficient fine-tuning

PEFT has proven to be an efficient strategy for fine-tuning a large, fundamental model for a specific usage (Zaken et al., 2021). Compared to fully fine-tuning, it keeps most parameters frozen and learns significantly fewer parameters, often less than 5% of the total. This enables efficient learning with faster updates. Studies have also shown that PEFT approaches work better than fully fine-tuning as they avoid catastrophic forgetting and generalize better to out-of-domain scenarios, especially in low-data regimes (Zaken et al., 2021). Among all PEFT strategies, Adaption (Hu et al., 2021; Zhang et al., 2023c; Yeh et al., 2023) stands out as an effective tool for fine-tuning large fundamental models for downstream tasks, not only in NLP but also in computer vision. Recent studies have shown that Adaption can be easily adopted in various downstream computer vision tasks (He et al., 2022; Chen et al., 2022).

However, existing adaptation methods still face challenges when the data modality of downstream tasks varies, i.e. 3D CT or MRI scans in the medical domain. In addition, there are significant differences in interactive behaviour between the source task and the downstream task, and the application of adaptation techniques to interactive models remains largely unexplored. Therefore, we believe Adaption is the most fitting technique for carrying SAM to the medical domain, but special designs are still needed.



a. ViT Block in SAM b. 2D Medical Image Adaption c. 3D Medical Image Adaption d. Decoder Adaption with Prompt

Fig. 1. Med-SA architecture. We use (b) as the encoder with a standard Adapter to process 2D medical images, and (c) incorporating SD-Trans to process 3D images. Then we use (d) as the decoder with HyP-Adpt to incorporate the prompts.

3. Method

3.1. Preliminary: SAM architecture

To begin with, we provide an overview of the SAM architecture. SAM comprises three main components: an image encoder, a prompt encoder, and a mask decoder. The image encoder is based on a standard Vision Transformer (ViT) pre-trained by MAE. Specifically, we use the ViT-H/16 variant, which employs 14 × 14 windowed attention and four equally-spaced global attention blocks, as shown in Fig. 1(a). The output of the image encoder is a 16× downsampled embedding of the input image. The prompt encoder can be either sparse (points, boxes) or dense (masks). In this paper, we focus only on the sparse encoder, which represents points and boxes as positional encodings summed with learned embeddings for each prompt type. The mask decoder is a Transformer decoder block modified to include a dynamic mask prediction head. The decoder uses two-way cross-attention to learn the interaction between the prompt and image embeddings. After that, SAM upsamples the image embedding, and an MLP maps the output token to a dynamic linear classifier, which predicts the target mask of the given image.

3.2. Med-SA architecture

Our objective is to enhance the medical capability of SAM for medical image segmentation tasks through fine-tuning. Rather than fully adjusting all parameters, we maintain the pre-trained SAM parameters frozen, devise an Adapter module, and integrate it into designated positions. The Adapter serves as a bottleneck model, consisting of a down-projection, ReLU activation, and up-projection sequentially, as illustrated in Fig. 1(b). The down-projection compresses the given embedding into a lower dimension using a simple MLP layer, while the up-projection expands the compressed embedding back to its original dimension using another MLP layer.

In the SAM encoder, we strategically place two adapters in each standard ViT block, as illustrated in Fig. 1(b). The first Adapter follows the multi-head attention, while the second is positioned in the residual path of the MLP layer following the multi-head attention. This architecture is intricately designed based on several insights and observations.

Firstly, multi-head attention plays a crucial role in ViT by enhancing the model's ability to capture diverse relationships in input sequences, allowing it to grasp both global and local dependencies in flattened image patches (Vaswani et al., 2017). This is why many adapter studies introduce adapters right after the multi-head attention blocks (Dettmers et al., 2023; Jie and Deng, 2023). However, recent studies (Dong et al., 2021; Chen et al., 2022) emphasize the importance of the MLP module in ViTs. It prevents ViTs from producing rank-1 matrices and avoids output degeneration. Building on this insight, we assert that a successful adapter design should focus not only on ViT multi-head self-attentions but also its MLPs. Compared to the fully-finetuned methods, this design is well-suited for the medical field, as the semantics of different organs and tissues vary significantly. The parallel structure allows for task-specific fine-tuning without compromising the general segmentation capability of SAM.

To achieve this, we integrate adapters into the MLP module parallelly, a strategy widely recognized for its effectiveness in feature ensemble (Wu et al., 2022b; Chen et al., 2022). The task-specific features from the adapter module complement the general features from the fixed branch, enriching the overall feature. We introduce a scaling factor s after the Adapter to balance these two types of features. An ablation study on adapter position and the scaling factor is presented in Section 4.7.

In the SAM decoder, we incorporate three adapters for each ViT block. The first Adapter is employed to integrate the prompt embedding, and to achieve this, we introduce a novel structure called the Hyper-Prompting Adapter (HyP-Adpt), which is further elaborated in Section 3.4.

The second Adapter in the decoder is deployed in exactly the same way as in the encoder, to adapt the MLP-enhanced embedding. The third Adapter is deployed after the residual connection of the image embedding-to-prompt cross-attention. Another residual connection and layer normalization are connected after the Adaption to output the final results.

3.3. SD-Trans

Adapting SAM to medical image segmentation poses a challenge due to the dimensional disparity between 2D images and the prevalent 3D modalities like MRI and CT scans. In clinical usage, understanding the correlation between slices is crucial for accurate decision-making. While SAM can be applied to each slice of a volume to obtain the final segmentation, it fails to consider the close volumetric correlation inherent in 3D medical image segmentation, as highlighted in previous studies (Hatamizadeh et al., 2022b,a). To address this limitation, we propose SD-Trans, inspired by image-to-video adaptation (Liu et al., 2019). The specific structure is depicted in Fig. 1(c).

As shown in the image, in each block, we bifurcate the attention operation into two branches: the space branch and the depth branch. For a given 3D sample with depth D, we input $D\times N\times L$ into the multi-head attention of the space branch, where N represents the number of embeddings, and L denotes the embedding length. Here, D corresponds to the number of operations, allowing the interaction to be applied over $N\times L$, capturing and abstracting spatial correlations as embeddings. In the depth branch, we transpose the input matrix to obtain $N\times D\times L$ and subsequently feed it into the same attention. Although using the same

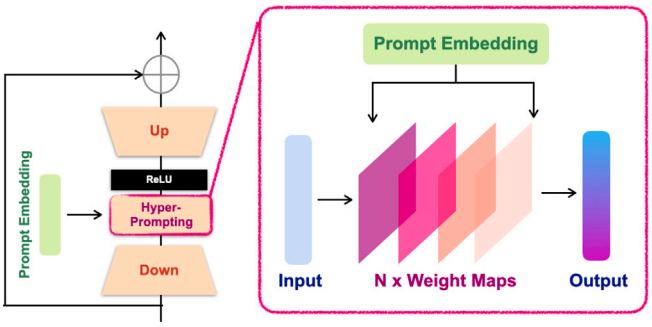


Fig. 2. HyP-Adpt architecture. We utilize Prompt Embedding to generate the weights applied to the Adapter Embedding.

attention mechanism, the interaction now occurs over $D \times L$, enabling the learning and abstraction of depth correlations. Finally, we transpose the results from the depth branch back to their original shape and add them to the space branch outputs, incorporating the depth information.

3.4. HyP-Adpt

While adaptation techniques have been applied to visual models in a few previous works, the application of adaptation to interactive visual models remains largely unexplored. The interactive behaviour between the natural scenes and the medical field exhibits significant differences. Therefore, it becomes crucial to incorporate the visual prompt, which plays a key role in the interactive model, into the Adapter. In this regard, we propose a solution called HyP-Adpt, aimed at achieving prompt-conditioned adaptation, as depicted in Fig. 1(d). The detailed structure of the Hyp-Adpt module is illustrated in Fig. 2.

The main idea of HyP-Adpt is to employ prompt embedding to generate weights for the Adapter for knowledge conditioning. Specifically, we utilize projection and reshaping operations to generate a sequence of weight maps from the prompt embedding. These weight maps are then directly applied (matrix product) to the adapter embedding. This approach enables wide and deep feature-level interaction while also significantly reducing the number of parameters required compared to generating an entire network. Formally, we conduct the hyper-prompting over the reduced embedding of the Adapter e^{down}. In the meantime, the prompt information (click location, click attribution, or bounding box location) is concatenated and reduced as prompt embedding e^{prompt}. Then we use e^{prompt} to generate the sequence of weights, taking one of them to illustrate would be:

$$W = Re(M(e^{prompt})), \tag{1}$$

where Re denotes reshape, and M denotes the MLP layer to project $e^{prompt} \in \mathcal{R}^{N \times L}$ to $e^{prompt} \in \mathcal{R}^{N \times (L^{in} * L^{out})}$, in which $*$ is value multiplication, L^{in} of the first weight will be the length of e^{down}, and L^{out} of the last weight will be the target length of the output. After that, we reshape e^{prompt} from 1D embedding to 2D weight $w^{prompt} \in \mathcal{R}^{N \times L^{in} \times L^{out}}$, and apply it over e^{down}, which can be represented as:

$$e^{down}_{n+1} = ReLU(Norm(e^{down}_{n} \otimes w^{prompt})), \tag{2}$$

where $\otimes$ is the matrix product. We normalize the elements along the length dimension and apply ReLU activation after that. We set 3 layers for the hyper-prompting, each weight is projected by individual MLP layers. HyP-Adpt helps to turn the parameter conditioned on the prompt information and be more flexible to different modalities and downstream tasks.

3.5. Prompt generation strategy

For interactive segmentation, we employ click prompts and bounding box (BBox) prompts during the model training process. To generate BBox prompts, we adopt the same approach as SAM. However, since the original SAM paper provides limited details on click prompt generation, we have devised our own and present here. Notably, we prompt each slice of 3D images to ensure segmentation consistency.

The fundamental concept behind our click prompt generation process involves using positive clicks to indicate foreground regions and negative clicks to indicate background regions. We combine random and iterative click-sampling strategies to train the model. We use random sampling for prompt initialization and then incorporate a few clicks using an iterative sampling procedure. This iterative sampling strategy emulates the interaction with a real user, as each new click is placed in the erroneous region of a prediction generated by the network using the set of previous clicks. We refer to Lin et al. (2020) for random sampling generation and Mahadevan et al. (2018) for simulating the iterative sampling process. The detailed implementation can be found in our released code.

4. Experiments

In this paper, we conduct extensive experiments to demonstrate the effectiveness of the proposed Med-SA. The main challenges in medical image segmentation lie in 3D image processing and the ambiguous boundaries between overlapped organs. Therefore, in Section 4.4 we choose the 3D abdominal multi-organ segmentation task to showcase the superb of Med-SA in the medical domain. To further verify the model's generalization ability, we perform comparative experiments on medical images of various modalities in Section 4.5. We also carry out ablation studies and detailed analysis of our method in Section 4.6 and Section 4.7

4.1. Dataset and evaluation metrics

We first leverage a large-scale and widely-used medical benchmark **BTCV** on 3D abdominal multi-organ segmentation for method evaluation (Fang and Yan, 2020). This dataset includes abdominal CT scans with 12 anatomical structures annotated by clinical experts. The examples are selected from 50 subjects with a total of 1463 axial contrast-enhanced abdominal clinical CT images. Each CT scan contains 85 to 198 slices with an axial size of 512 × 512.

We further utilize datasets for different tasks with various modalities for comparison, to validate the robustness and generalization capability of our method, including:

REFUGE2 (Fang et al., 2022): This is a 2D retinal fundus image dataset for 2 segmentation tasks (optic disc and optic cup), which contains 1200 RGB images at a resolution of 2124 × 2056 annotated by experts.

BraTS2021 (Baid et al., 2021): This is a 3D dataset for brain glioblastoma sub-region segmentation, which contains 1280 multi-parametric MRI scans from multiple institutions. Each MRI scan contains 155 slices with an axial size of 240 × 240.

TNMIX: This is a widely-used benchmark for 2D thyroid nodule segmentation in ultrasound images of different resolutions, which is a mixed dataset containing 4554 examples from TNSCUI (Ma et al., 2017) and 637 examples from DDTI (Pedraza et al., 2015).

ISIC2019 (Milton, 2019): This is a 2D dermoscopic image dataset for skin lesion analysis and contains 25331 images with human-annotated labels for melanoma segmentation. The data come from various centres and have different resolutions.

These datasets cover 17 segmentation subjects altogether, which are all publicly available and have been widely used as benchmarks for evaluating medical image segmentation. We take their default splitting of training, validation, and test set. For model evaluation, we use the Dice score and Hausdorff Distance (HD) as key metrics to assess pixel-wise segmentation accuracy and segmentation boundary quality, respectively. Additionally, we report the mean intersection-over-union (mIoU) for several segmentation tasks to facilitate better comparisons.

Table 1
The comparison of Med-SA with SOTA methods over the BTCV dataset evaluated by Dice Score (%) and average HD95. Best results are denoted in **bold**.

Model	Prms (M)	Tunable Prms (M)	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Veins	Panc.	AG	Avg	HD95 ↓
TransUNet	37	37	95.2	92.7	92.9	66.2	75.7	96.9	88.9	92.0	83.3	79.1	77.5	63.7	83.8	8.12
EnsDiff	32	32	93.8	93.1	92.4	77.2	77.1	96.7	91.0	86.9	85.1	80.2	77.1	74.5	85.4	6.62
SegDiff	32	32	95.4	93.2	92.6	73.8	76.3	95.3	92.7	84.6	83.3	79.6	78.2	72.3	84.7	6.73
UNetr	104	104	96.8	92.4	94.1	75.0	76.6	97.1	91.3	89.0	84.7	78.8	76.7	74.1	85.6	6.35
Swin-UNetr	138	138	97.1	93.6	94.3	79.4	77.3	97.5	92.1	89.2	85.3	81.2	79.4	76.5	86.9	5.85
nnUNet	16	16	94.2	89.4	91.0	70.4	72.3	94.8	82.4	87.7	78.2	72.0	68.0	61.6	80.2	23.49
CoTr	41.9	41.9	96.3	88.5	90.8	66.6	78.0	97.1	88.2	91.2	88.0	78.1	83.1	74.1	85.0	6.12
TransUNet3D	41.4	41.4	92.4	88.9	85.7	82.0	75.5	97.2	85.1	93.0	86.8	82.1	82.9	75.5	85.6	6.44
Med-SA everything	636	13	97.6	91.3	96.8	82.1	81.7	98.4	93.0	91.0	87.3	80.6	76.1	81.3	88.1	3.38
Med-SA 1 point	636	13	97.8	93.5	96.6	82.3	81.8	98.1	93.1	91.5	87.7	81.1	76.7	80.9	88.3	3.22
Med-SA 3 points	636	13	98.0	93.6	96.8	82.6	82.1	98.6	93.4	91.7	87.8	81.3	77.1	81.8	88.7	3.06
Med-SA BBox 0.5	636	13	95.4	91.0	95.2	81.0	80.7	97.5	92.8	91.2	86.8	80.9	76.9	81.3	87.6	4.78
Med-SA BBox 0.75	636	13	98.5	94.7	97.5	84.2	80.8	98.3	94.2	93.9	89.9	85.2	79.0	82.3	**89.8**	**2.18**

4.2. Implementation details

In this study, we implement the Med-SA pipeline primarily following the official SAM (VIT-H as the backbone). For 2D medical image training, we adhere to the SAM's default training settings. For 3D medical image training, we use a smaller batch size of 16. For the REFUGE2, TNMIX, and ISIC datasets, we train the model for 40 epochs. For the 3D BTCV and BraTS datasets, we extend the training to 60 epochs. All the experiments are implemented with the PyTorch platform and trained/tested on 4 NVIDIA A100 GPUs. We utilize the default settings to reproduce the comparison methods. For all datasets, 2D images are resized to have a long side of 1024 pixels, with padding applied to create a 1024×1024 pixel square. For 3D images, the original volumes are resized to a $128 \times 128 \times 128$ scale without cropping or padding following the common setting (Isensee et al., 2021).

As for the prompt for the interactive models, we conduct four different prompt settings following previous works, including: (1) a random 1 positive point, denoted as "1-point", (2) three positive points, denoted as "3-points", (3) bounding boxes with 50% overlapping of the target, denoted as "BBox 0.5", and (4) bounding boxes with 75% overlapping of the target, denoted as "BBox 0.75".

Notably, to ensure segmentation consistency, prompts were provided separately for each slice of the 3D images. To minimize the randomness introduced by random prompts, we repeated the Med-SA experiments five times across all datasets. Due to space constraints, the tables in this section present the average results, while variances for specific tasks are discussed in Section 5.

4.3. Comparison methods

In this paper, the comparison methods in all experiments can be categorized into the following four types:

- **Well-recognized methods**: These methods have demonstrated excellent performance across various tasks (both 2D and 3D) in medical image segmentation and are widely used in the domain, including nnUNet (Isensee et al., 2021), TransUNet (Chen et al., 2021a), UNetr (Hatamizadeh et al., 2022b), Swin-UNetr (Hatamizadeh et al., 2022a), CoTr (Xie et al., 2021), TransUNet3D (Chen et al., 2024), EnsDiff (Wolleb et al., 2021), and SegDiff (Amit et al., 2021).
- **Interactive segmentation methods**: including vanilla SAM, fully fine-tuned MedSAM (Ma and Wang, 2023), SAMed (Zhang et al., 2023b), SAM-Med2D (Cheng et al., 2023), SAM-U (Deng et al., 2023a), AdaptiveSAM (Paranjape et al., 2024), VMN (Zhou et al., 2023), and FCFI (Wei et al., 2023). In FCFI, we use ConvNext-v2-H (Woo et al., 2023) as the backbone.
- **3D SAM-based methods**: These methods are designed to leverage the SAM for 3D image segmentation, including SAM3D (Bui et al., 2023), 3DSAM-adt (Gong et al., 2023), MA-SAM (Chen et al., 2023), Promise (Li et al., 2023), and SAM-Med3D (Wang et al., 2023).
- **Task-specific methods**: These methods are designed for particular diseases, including ResUnet (Yu et al., 2019) and BEAL (Wang et al., 2019) for optic cup segmentation, TransBTS (Wang et al., 2021b) and EnsemDiff (Wolleb et al., 2021) for brain tumor segmentation, MTSeg (Gong et al., 2021) and UltraUNet (Chu et al., 2021) for thyroid nodule segmentation, and FAT-Net (Wu et al., 2022a) and BAT (Wang et al., 2021a) for melanoma segmentation.

Notably, we have tried our best to ensure our comparative analysis is ***extensive and comprehensive***. We have carefully considered multiple aspects as shown above and have diligently gathered the current SOTA methods to ensure a thorough evaluation. Considering that our Med-SA is one of the first works of adapting SAM to the medical domain, which received high interest and many following methods are designed based on our approach (Yue et al., 2024; Song et al., 2024; Cheng et al., 2024; Deng et al., 2024; Yan et al., 2024), we do not include these preprints for comparison.

4.4. Comparison on abdominal multi-organ segmentation

To verify the general performance of our proposed model, we compare it with SOTA segmentation methods on the multi-organ segmentation dataset BTCV. Table 1 shows the quantitative results comparing to well-recognized methods, while Table 2 presents the comparison results with interactive segmentation methods and 3D SAM-based methods.

In Table 1, we can see that Med-SA achieves a significant improvement over existing prevailed methods even with only a 1-point prompt. Remarkably, on the BTCV dataset, the 1-point Med-SA outperforms the previous SOTA for the average of all 12 organs and achieves the best performance in 8 organs. As we provide more fine-grained prompts, the results continue to improve, reaching a final Dice of 89.8% and HD95 of 2.18 with BBox 0.75. This result outperforms the previous SOTA (Swin-UNetr) by a significant margin of 2.9% and 3.67, when Swin-UNetr consists of 138M tunable parameters and we only update 13M parameters. Notably, our Med-SA also outperforms CoTr and TransUnet3D by 4.8% and 4.2% on Dice score even if they are specifically designed for 3D images. In addition, we also perform "segment everything" automatically following the vanilla SAM setting, yielding an average Dice of 88.1%, which is slightly worse than its 1-point prompt performance but still outperforms Swin-UNetr by 1.2%.

Surprisingly, Med-SA still leads in the performance among interactive models and 3D SAM-based methods as shown in Table 2, even outperforming fully fine-tuned models (e.g., MedSAM Ma and Wang,

Table 2
The comparison of Med-SA with other methods and 3D-based SAM methods over the BTCV dataset evaluated by Dice Score (%) and average HD95.

Model	Training time (H)	Prms (M)	Tunable Prms (M)	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Veins	Panc.	AG	Avg	HD95↓
SAM 1 points	0	636	0	51.8	68.6	79.1	54.3	58.4	46.1	56.2	61.2	40.2	55.3	51.1	35.4	54.8	55.6
SAM 3 points	0	636	0	62.2	71.0	81.2	61.4	60.5	51.3	67.3	64.5	48.3	62.8	56.4	39.5	63.1	48.1
SAM BBox 0.5	0	636	0	34.6	58.5	59.2	37.5	42.6	37.7	45.1	53.6	39.2	57.6	42.6	20.2	44.0	65.6
SAM BBox 0.75	0	636	0	41.5	62.1	67.8	58.0	59.5	46.9	52.1	61.2	53.9	65.5	58.8	32.7	55.0	55.2
SAMed 1 points	1.25	636	21	86.2	71.0	79.8	67.7	73.5	94.4	76.6	87.4	79.8	77.5	57.9	79.0	77.6	31.7
SAMed 3 points	1.25	636	21	88.5	73.1	81.3	68.0	75.5	94.7	77.3	89.1	81.1	78.3	62.1	79.5	79.0	26.8
SAMed BBox 0.5	1.25	636	21	72.2	65.7	71.6	64.4	61.7	77.5	62.1	68.0	71.3	50.1	58.2	54.8	64.5	43.7
SAMed BBox 0.75	1.25	636	21	87.1	74.1	82.5	69.5	73.2	93.8	74.6	87.2	85.0	76.1	66.0	77.2	78.9	28.3
SAM-Med2D 1 points	3.42	636	636	87.3	88.4	93.2	79.5	79.0	94.3	88.9	87.2	79.6	81.3	77.9	79.7	84.7	10.2
SAM-Med2D 3 points	3.42	636	636	87.7	88.5	93.6	79.8	79.1	94.6	89.1	87.8	79.6	81.5	78.2	79.9	84.9	9.6
SAM-Med2D BBox 0.5	3.42	636	636	86.7	88.1	92.8	79.1	78.6	94.0	88.5	86.8	79.0	82.7	76.5	79.1	84.3	11.3
SAM-Med2D BBox 0.75	3.42	636	636	89.1	89.8	92.5	81.8	80.1	95.2	91.1	87.5	83.4	81.8	79.0	78.7	85.8	9.4
SAM-U 1 points	5.51	636	636	86.8	77.6	83.4	69.0	71.0	92.2	80.5	86.3	84.4	78.2	61.1	78.0	79.0	22.2
SAM-U 3 points	5.51	636	636	87.9	79.4	84.3	70.8	73.1	93.0	81.2	87.0	85.5	78.8	62.3	78.5	80.2	18.9
SAM-U BBox 0.5	5.51	636	636	81.5	74.0	78.1	67.8	66.0	74.5	77.6	74.9	70.6	63.4	58.9	71.3	71.5	38.7
SAM-U BBox 0.75	5.51	636	636	91.0	81.2	86.0	74.3	75.0	94.1	83.4	89.2	85.8	79.1	64.2	78.7	81.8	17.5
VMN 1 points	3.80	58	58	80.3	78.8	80.1	78.3	71.2	87.0	82.1	83.2	82.5	74.2	65.5	71.0	77.9	32.8
VMN 3 points	3.80	58	58	83.4	81.2	85.0	82.2	75.9	88.0	83.7	85.0	84.4	76.2	67.3	73.1	80.5	21.6
VMN BBox 0.5	3.80	58	58	81.2	76.5	78.0	77.6	72.9	82.3	78.5	75.8	77.3	76.5	66.3	73.1	76.3	35.0
VMN BBox 0.75	3.80	58	58	84.6	80.9	85.4	85.8	76.7	85.2	84.9	85.5	83.1	79.0	68.1	74.2	81.1	19.2
FCFI 1 points	7.21	718	718	87.6	83.4	88.9	79.5	78.1	94.5	88.7	92.1	89.7	82.9	78.0	76.0	85.8	9.8
FCFI 3 points	7.21	718	718	88.0	84.1	89.3	80.1	79.5	95.6	89.2	92.8	90.6	83.7	84.0	78.8	86.2	9.2
FCFI BBox 0.5	7.21	718	718	84.1	81.9	89.5	81.2	78.3	84.3	83.9	86.2	84.0	79.5	71.1	73.0	81.4	16.9
FCFI BBox 0.75	7.21	718	718	94.8	92.9	93.0	80.6	89.7	90.1	92.0	81.8	83.3	79.7	77.5	79.1	86.2	9.6
AdaptiveSAM 1 points	0.87	636	16	86.0	69.6	94.7	86.5	71.3	83.3	92.7	47.0	87.2	92.9	38.2	66.2	76.3	33.4
AdaptiveSAM 3 points	0.87	636	16	65.9	61.0	94.7	93.3	60.0	73.2	80.7	82.2	93.1	75.9	47.1	91.7	76.6	33.1
AdaptiveSAM BBox 0.5	0.87	636	16	91.3	55.7	62.7	75.8	69.8	38.7	87.6	92.6	85.4	93.0	93.8	66.2	76.1	31.5
AdaptiveSAM BBox 0.75	0.87	636	16	75.0	94.9	62.0	83.2	71.4	80.5	91.1	71.4	35.7	82.7	88.2	89.8	77.2	30.6
SAM3D 1 points	11.08	636	636	62.2	82.2	65.5	78.9	57.7	77.2	60.1	62.4	80.7	89.2	64.6	77.2	71.5	38.2
SAM3D 3 points	11.08	636	636	80.7	80.5	31.1	87.3	79.2	95.0	72.9	84.5	55.2	85.4	58.9	81.8	74.4	34.3
SAM3D BBox 0.5	11.08	636	636	69.8	94.4	56.4	50.6	84.2	89.6	84.2	93.9	54.0	66.9	32.5	94.7	72.6	36.0
SAM3D BBox 0.75	11.08	636	636	58.4	79.6	91.0	81.1	86.4	64.1	51.1	78.5	86.3	63.7	90.9	90.0	76.8	33.3
3DSAM-adt 1 points	1.16	636	21	89.4	73.3	44.9	52.4	82.9	81.2	77.3	44.7	82.3	81.7	70.1	69.3	70.8	41.2
3DSAM-adt 3 points	1.16	636	21	88.5	86.7	74.8	46.6	52.3	57.9	81.3	81.1	80.8	92.5	75.6	69.9	74.0	35.7
3DSAM-adt BBox 0.5	1.16	636	21	70.7	68.9	37.6	78.0	93.7	54.5	68.6	67.5	83.4	75.8	74.2	86.8	71.6	37.4
3DSAM-adt BBox 0.75	1.16	636	21	71.1	72.6	94.0	74.1	87.5	85.2	66.5	68.6	81.6	89.9	75.4	46.1	76.1	31.8
MA-SAM 1 points	8.24	636	636	42.6	92.5	66.8	91.9	79.3	78.5	55.4	30.4	87.8	67.9	78.4	88.9	71.7	37.3
MA-SAM 3 points	8.24	636	636	69.7	79.4	70.9	68.9	75.8	94.7	92.7	69.5	43.4	68.9	70.7	63.2	72.3	36.6
MA-SAM BBox 0.5	8.24	636	636	45.4	92.0	68.2	65.1	84.1	70.4	90.0	60.8	33.4	68.5	91.8	71.1	70.1	39.0
MA-SAM BBox 0.75	8.24	636	636	54.3	70.3	66.0	73.4	75.8	42.2	56.4	92.3	93.6	71.7	85.8	93.2	72.9	36.1
Promise 1 points	7.83	647	647	71.2	93.3	71.2	86.8	88.6	90.4	92.5	77.8	86.7	92.1	80.7	92.8	85.3	7.1
Promise 3 points	7.83	647	647	71.9	90.8	91.6	82.8	94.2	81.8	89.7	78.8	91.9	86.6	91.4	92.9	86.4	6.9
Promise BBox 0.5	7.83	647	647	93.5	86.2	94.1	80.7	80.5	92.3	83.5	56.3	84.2	88.6	92.8	87.6	85.0	7.6
Promise BBox 0.75	7.83	647	647	89.3	94.0	92.5	78.3	76.4	77.5	91.2	91.4	87.9	83.9	94.5	92.6	87.5	6.4
SAM-Med3D 1 points	9.55	677	677	69.9	83.4	86.7	91.8	88.2	85.1	86.7	86.8	85.7	93.2	74.1	90.7	85.2	7.8
SAM-Med3D 3 points	9.55	677	677	91.3	79.1	86.2	90.0	82.6	87.4	94.0	91.0	91.8	91.2	70.9	68.3	85.3	7.4
SAM-Med3D BBox 0.5	9.55	677	677	83.9	81.8	86.6	94.9	93.3	69.2	90.6	91.6	78.2	76.3	95.0	79.1	85.0	7.7
SAM-Med3D BBox 0.75	9.55	677	677	93.0	84.9	91.6	67.3	81.8	88.0	90.6	78.3	93.4	78.5	88.2	91.5	85.6	7.0
MedSAM 1 point	2.83	636	636	75.1	81.4	88.5	76.6	72.1	90.1	85.5	87.2	74.6	77.1	76.0	70.5	80.3	17.5
MedSAM 3 points	2.83	636	636	75.8	83.1	88.9	78.2	73.3	91.7	85.8	87.6	75.5	77.6	76.3	71.6	82.0	15.8
MedSAM BBox 0.5	2.83	636	636	62.1	73.6	80.1	72.1	71.5	81.1	71.4	77.0	62.2	61.8	63.0	54.5	69.2	42.4
MedSAM BBox 0.75	2.83	636	636	74.6	84.2	87.3	77.2	74.5	89.7	86.0	88.9	74.3	74.5	73.9	70.1	80.4	18.6
Med-SA everything	0.37	636	13	97.6	91.3	96.8	82.1	81.7	98.4	93.0	91.0	87.3	80.6	76.1	81.3	88.1	3.4
Med-SA 1 point	0.37	636	13	97.8	93.5	96.6	82.3	81.8	98.1	93.1	91.5	87.7	81.1	76.7	80.9	88.3	3.2
Med-SA 3 points	0.37	636	13	98.0	93.6	96.8	82.6	82.1	98.6	93.4	91.7	87.8	81.3	77.1	81.8	88.7	3.1
Med-SA BBox 0.5	0.37	636	13	95.4	91.0	95.2	81.0	80.7	97.5	92.8	91.2	86.8	80.9	76.9	81.3	87.6	4.8
Med-SA BBox 0.75	0.37	636	13	98.5	94.7	97.5	84.2	80.8	98.3	94.2	93.9	89.9	85.2	79.0	82.3	**89.8**	**2.2**

2023) across all prompt variations. With the proposed SD-Trans and HyP-Adpt, Med-SA outperforms MedSAM by updating only 2% of its total tunable parameters (13M v.s. 636M), which highlights the effectiveness of the proposed techniques. It can also be observed that SAM's zero-shot performance is generally inferior to fully-trained models in medical image segmentation tasks, regardless of the prompt used. While this comparison may seem unfair, SAM has demonstrated superior zero-shot performance in natural image datasets, which indicates that SAM's zero-shot transferability is less effective for medical images This finding has also been observed in previous studies (Deng et al., 2023b; Roy et al., 2023), which emphasizes the need for specific techniques to adapt SAM to medical domain.

When comparing the performance of different prompts in interactive segmentation models, we notice that 3-points slightly outperform 1-point prompts. BBox 0.75 often performs comparably or better than 3-point prompts. However, it is important to note that BBox 0.5 yields subpar performance, indicating the significance of accurate bounding box annotations for achieving performance improvements. All interactive models exhibit similar behaviour across different prompts, demonstrating consistency in their response to prompts.

Table 3
The comparison of Med-SA with SAM and SOTA segmentation methods on different image modalities. The grey background denotes the methods proposed for that/those particular tasks.

Model	Optic-Disc		Optic-Cup		Brain-Tumor			Thyroid Nodule		Melanoma	
	Dice	IoU	Dice	IoU	Dice	IoU	HD95↓	Dice	IoU	Dice	IoU
ResUNet	92.9	85.5	80.1	72.3	78.4	71.3	18.71	78.3	70.7	87.1	78.2
BEAL	93.7	86.1	83.5	74.1	78.8	71.7	18.53	78.6	71.6	86.6	78.0
TransBTS	94.1	87.2	85.4	75.7	87.6	78.44	12.44	83.8	75.5	88.1	80.6
EnsemDiff	94.3	87.8	84.2	74.4	88.7	80.9	10.85	83.9	75.3	88.2	80.7
MTSeg	90.3	83.6	82.3	73.1	82.2	74.5	15.74	82.3	75.2	87.5	79.7
UltraUNet	91.5	82.8	83.1	73.8	84.5	76.3	14.03	84.5	76.2	89.0	81.8
FAT-Net	91.8	84.8	80.9	71.5	79.2	72.8	17.35	80.8	73.4	90.7	83.9
BAT	92.3	85.8	82.0	73.2	79.6	73.5	15.49	81.7	74.2	91.2	84.3
SegDiff	92.6	85.2	82.5	71.9	85.7	77.0	14.31	81.9	74.8	87.3	79.4
nnUNet	94.7	87.3	84.9	75.1	88.5	80.6	11.20	84.2	76.2	90.8	83.6
TransUNet	95.0	87.7	85.6	75.9	86.6	79.0	13.74	83.5	75.1	89.4	82.2
UNetr	94.9	87.5	83.2	73.3	87.3	80.6	12.81	81.7	73.5	89.7	82.8
Swin-UNetr	95.3	87.9	84.3	74.5	88.4	81.8	11.36	83.5	74.8	90.2	83.1
CoTr	95.8	87.1	85.9	75.3	87.7	81.2	12.91	82.6	74.5	90.6	83.9
TransUNet3D	-	-	-	-	88.9	81.7	11.08	-	-	-	-
Med-SA everything	97.1	89.0	85.9	76.3	88.9	81.9	10.97	85.8	77.2	92.2	83.9
Med-SA 1 point	97.4	89.5	86.8	78.8	89.1	82.0	10.38	86.3	78.7	92.6	84.1
Med-SA 3 points	97.9	89.8	87.1	79.0	89.8	82.3	10.11	86.7	79.4	**93.4**	**84.7**
Med-SA BBox 0.5	97.6	89.6	86.4	78.5	89.5	81.9	10.35	86.6	78.9	92.1	83.0
Med-SA BBox 0.75	**98.3**	**90.1**	**87.5**	**79.9**	**90.5**	**83.0**	**9.50**	**88.4**	**80.4**	93.0	84.2

Table 4
The comparison of Med-SA with SAM and other interactive segmentation methods on different image modalities. Performance is omitted (–) if the algorithm fails over 70% of the samples.

Model	Optic-Disc		Optic-Cup		Brain-Tumor			Thyroid Nodule		Melanoma	
	Dice	IoU	Dice	IoU	Dice	IoU	HD95↓	Dice	IoU	Dice	IoU
SAM 1 points	–	–	–	–	63.2	47.6	32.53	–	–	81.6	70.4
SAM 3 points	–	–	–	–	71.3	64.5	28.74	–	–	85.8	77.5
SAM BBox 0.5	–	–	–	–	51.2	44.6	38.56	–	–	75.3	64.8
SAM BBox 0.75	–	–	–	–	74.6	62.1	27.51	–	–	85.7	74.4
SAMed 1 point	89.9	81.8	80.7	70.8	77.3	68.8	19.07	78.9	71.2	87.4	78.9
SAMed 3 points	91.2	82.3	81.5	71.6	77.8	69.0	18.76	79.8	72.5	88.0	79.3
SAMed BBox 0.5	90.3	81.3	81.0	70.1	76.1	66.8	21.12	77.4	69.5	86.0	77.2
SAMed BBox 0.75	91.8	82.7	82.4	72.5	78.1	70.3	17.86	79.1	72.7	87.1	78.6
SAM-Med2D 1 point	92.1	83.7	82.0	75.3	82.9	74.1	16.20	80.3	73.6	87.8	78.3
SAM-Med2D 3 points	93.5	84.2	83.1	76.1	84.5	75.8	14.81	81.2	74.5	88.1	79.0
SAM-Med2D BBox 0.5	91.2	81.8	81.7	74.9	79.3	72.2	18.78	79.1	71.3	87.1	77.8
SAM-Med2D BBox 0.75	94.8	87.0	83.7	76.6	85.2	76.7	14.29	82.7	76.1	89.0	80.6
SAM-U 1 point	91.2	82.4	81.5	73.2	81.0	72.9	17.26	79.8	74.0	88.7	79.6
SAM-U 3 points	92.8	83.6	82.7	74.6	82.7	74.3	16.18	82.1	74.6	89.2	80.3
SAM-U BBox 0.5	90.8	81.5	81.9	75.1	78.9	71.8	17.53	77.6	70.2	87.8	78.1
SAM-U BBox 0.75	95.2	88.6	83.0	74.9	81.8	74.0	16.46	83.9	75.8	90.6	81.8
VMN 1 point	92.5	83.9	82.8	76.1	83.4	74.6	17.13	81.4	74.2	88.3	79.1
VMN 3 points	94.2	86.1	83.6	76.9	84.4	75.8	16.79	81.8	75.1	89.0	80.6
VMN BBox 0.5	91.4	82.6	82.0	71.3	80.2	72.0	16.96	80.2	73.4	87.8	79.0
VMN BBox 0.75	95.0	87.8	84.2	77.1	85.9	77.0	14.32	82.4	75.5	90.4	81.6
FCFI 1 point	95.5	88.3	85.7	77.7	87.0	78.2	12.32	84.3	75.7	89.6	81.5
FCFI 3 points	96.2	88.7	86.1	78.3	88.5	79.9	11.60	85.8	76.8	91.4	82.8
FCFI BBox 0.5	94.6	85.3	84.3	77.4	85.1	76.6	14.33	82.5	74.8	89.0	80.4
FCFI BBox 0.75	97.6	89.8	87.0	78.6	89.0	81.1	11.20	86.0	78.2	91.9	83.7
3DSAM-adt 1 points	90.7	82.1	80.8	71.0	86.1	78.4	12.86	79.0	72.4	88.2	79.1
3DSAM-adt 3 points	92.1	84.3	82.1	73.5	88.0	80.0	11.12	80.8	72.9	89.6	80.8
3DSAM-adt BBox 0.5	90.3	81.7	80.2	72.3	85.3	77.9	12.92	78.8	72.0	87.8	78.5
3DSAM-adt BBox 0.75	93.8	85.7	83.6	74.1	89.5	81.7	11.08	81.6	73.7	91.2	82.0
SAM-Med3D 1 points	91.8	82.6	81.7	74.0	88.2	79.1	11.84	81.5	74.2	88.2	78.8
SAM-Med3D 3 points	92.3	82.8	82.1	75.6	88.7	79.8	11.32	81.8	74.9	88.7	79.4
SAM-Med3D BBox 0.5	91.1	82.0	80.8	73.5	87.3	78.1	12.24	80.8	73.0	87.1	77.2
SAM-Med3D BBox 0.75	92.8	83.4	84.5	76.8	89.8	81.0	10.64	83.7	76.2	90.2	81.2
MedSAM 1 point	92.9	85.5	82.1	73.8	81.5	74.3	15.68	81.3	74.7	86.8	77.5
MedSAM 3 points	93.8	86.2	82.8	74.2	82.3	74.8	15.19	81.6	75.1	87.5	78.6
MedSAM BBox 0.5	92.6	85.3	82.0	75.2	82.0	74.7	15.05	82.4	75.5	88.5	79.2
MedSAM BBox 0.75	94.6	86.7	82.8	75.9	83.6	75.6	14.90	82.8	75.7	88.9	79.8
Med-SA everything	97.1	89.2	86.2	78.5	88.7	80.3	10.92	85.4	77.5	91.8	83.0
Med-SA 1 point	97.4	89.5	86.8	78.8	89.1	81.8	10.38	86.3	78.7	92.6	84.1
Med-SA 3 points	97.9	89.8	87.1	79.0	89.8	82.3	10.11	86.7	79.4	**93.4**	**84.7**
Med-SA BBox 0.5	97.6	89.6	86.4	78.5	89.5	81.9	10.35	86.6	78.9	92.1	83.0
Med-SA BBox 0.75	**98.3**	**90.1**	**87.5**	**79.9**	**90.5**	**83.0**	**9.50**	**88.4**	**80.4**	93.0	84.2

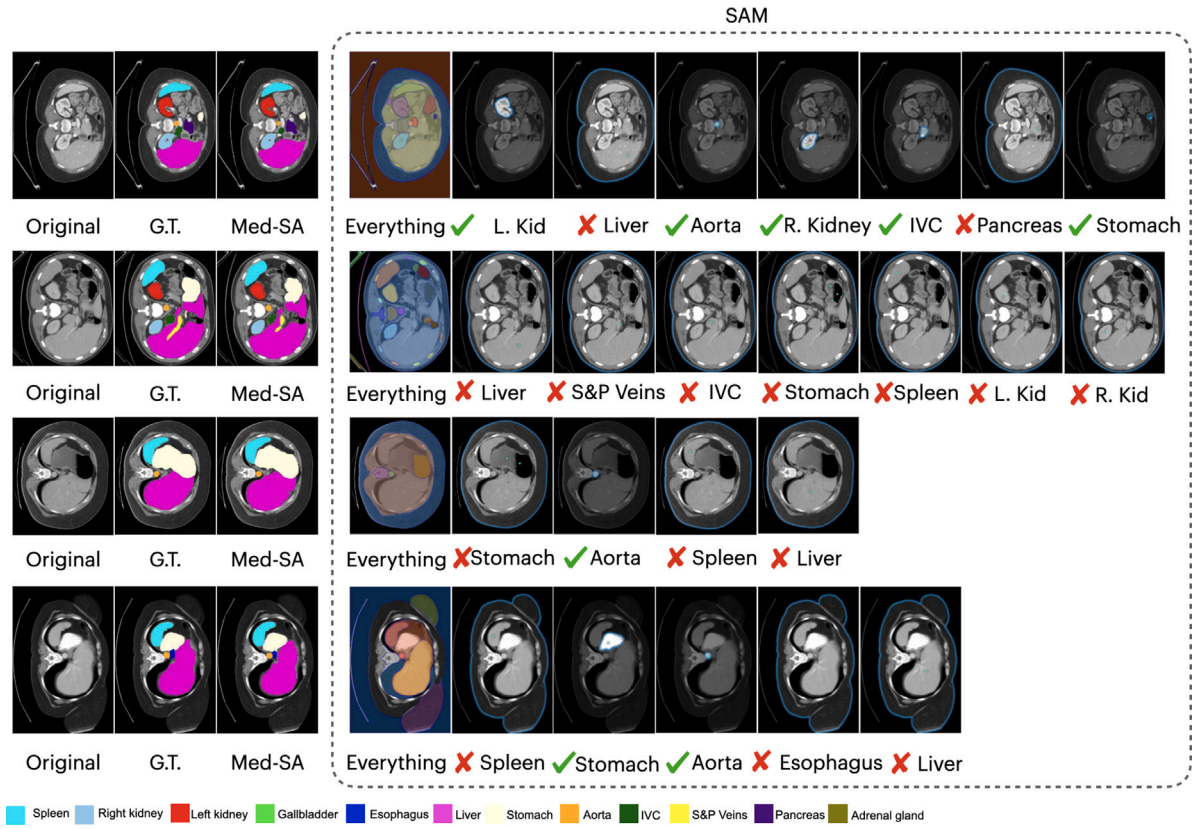


Fig. 3. Visual comparison of Med-SA and SAM on abdominal multi-organ segmentation. We use a **Checkmark** to represent SAM correctly found the organ and a **Cross** to represent it lost.

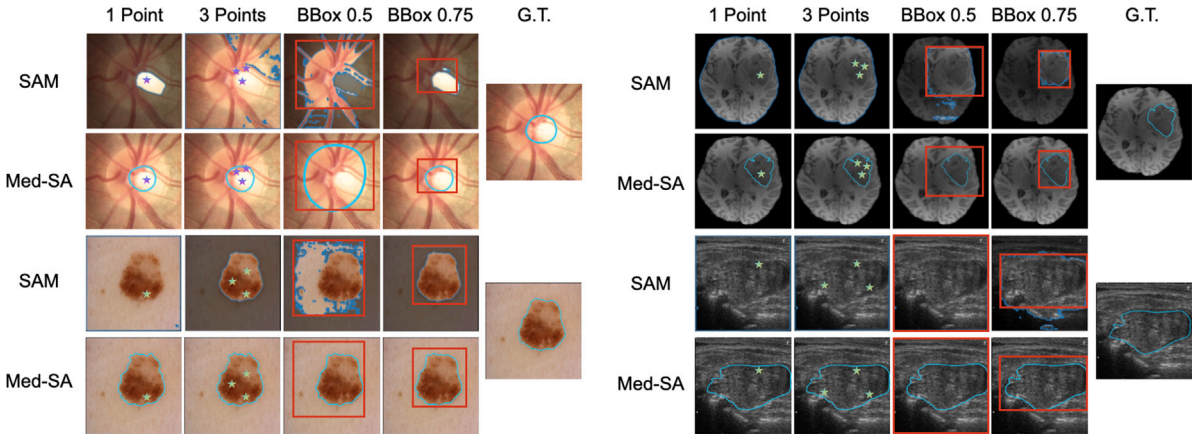


Fig. 4. Visual comparison of Med-SA and SAM on medical image segmentation with four different modalities. Top-left: optic disc and cup segmentation from the fundus image. Top-right: Brain tumor segmentation from the Brain MRI. Bottom-left: melanoma segmentation from the dermoscopic image. Bottom-right: thyroid nodule segmentation from the ultrasound image.

Fig. 3 presents a qualitative comparison of the performance between Med-SA and SAM. The figure shows that Med-SA segments accurately on parts that are difficult to recognize by the human eye. Conversely, SAM fails in many cases where the organ boundaries are visually clear. This further underscores the necessity of fine-tuning a general segmentation model on medical images to achieve optimal performance.

4.5. Generalizing to more tasks with various modalities

In this section, we further compare Med-SA to SOTA methods across four different medical image segmentation tasks involving various modalities of both 2D and 3D medical imaging. Table 3 shows the quantitative results comparing to well-recognized methods and task-specific methods, while Table 4 presents the comparison results with interactive segmentation methods and some 3D SAM-based methods. The performance is evaluated using the Dice score, IoU, and HD95 metrics.

From Table 3, we can see that specifically optimized methods often perform well within their respective domains but experience drops in performance in other domains. For example, UltraUNet achieves the previous SOTA for thyroid nodule segmentation but performs the worst in optic disc segmentation compared with others. On the other hand, general methods often achieve good results across most modalities but fail to outperform specialized methods in specific tasks such as

Table 5
An ablation study on SD-Trans and HyP-Adpt.

2D-3D	Prompt-Condition			BTCV	Optic-Disc	Optic-Cup	BrainTumor	ThyroidNodule	Melanoma
SD-Trans	Add	Concat	HyP-Adpt	Ave-Dice (%)	Dice (%)	Dice (%)	Dice (%)	Dice (%)	Dice (%)
				79.3	90.1	80.1	77.5	76.5	89.2
✓				84.7	–	–	81.7	–	–
✓	✓			86.1	94.6	83.4	83.9	83.7	93.8
✓		✓		86.4	95.7	84.0	85.1	84.8	94.5
✓			✓	**88.3**	**97.4**	**86.8**	**87.6**	**86.3**	**96.3**

Table 6
Results with different adapter positions.

Main-Branch	MLP-Adapt	BTCV	Optic-Disc	Optic-Cup	BrainTumor	ThyroidNodule	Melanoma
Adapter		Ave-Dice (%)	Dice (%)	Dice (%)	Dice (%)	Dice (%)	Dice (%)
✓		87.4	96.8	85.4	87.1	85.6	94.9
	✓	87.1	96.2	85.7	87.2	85.4	95.5
✓	✓	**88.3**	**97.4**	**86.8**	**87.6**	**86.3**	**96.3**

Table 7
Ablation on different hyper-parameters settings.

(a) Adapter ranks			(b) Adapter inserted layers			(c) Scaling factor s		
Ranks	#Params	BTCV (Dice (%))	Layers	#Params	BTCV (Dice (%))	s	BTCV (Dice (%))	Optic-Disc (Dice (%))
1	1.58M	75.2	1–16	6.55M	77.3	0.01	85.0	96.1
16	4.33M	85.6	17–32	6.55M	86.1	0.05	**88.3**	**97.4**
32	7.19M	86.9	1–32	13.10M	**88.3**	0.10	87.7	96.9
64	13.10M	**88.3**				0.20	87.2	96.6
256	48.67M	87.8						

brain tumor and melanoma segmentation. Meanwhile, Med-SA shows SOTA performance on all segmentation tasks, even surpassing methods specifically designed for 3D images, such as EnsemDiff and TransUNet3D in brain tumor segmentation. This further highlights the strong generalizability of our Med-SA.

Turning our attention to the interactive models as shown in Table 4, it can be found that on 2D images, zero-shot SAM struggles with organs/tissues that have ambiguous boundaries in medical images, such as optic cup and thyroid nodule segmentation. PEFT-based SAMed also shows inferior performance. Our Med-SA achieves SOTA performance across all segmentation tasks while using much fewer parameters. On 3D images, other methods either perform poorly or rely heavily on prompt accuracy (FCFI), 3DSAM-adt and SAM-Med3D demonstrate significant efficacy in 3D medical image segmentation through multi-layer and slice-by-slice aggregation, while Med-SA still outperforms the previous SOTA SAM-Med3D by 0.7% in Dice score, 2.0% in IoU, and 1.64 in HD95 metric, and remains stable under different prompts. This demonstrates Med-SA's great generalization ability when facing segmentation tasks in different modalities, as well as its stability regarding prompts during 3D image processing.

Fig. 4 provides a qualitative comparison between our proposed Med-SA model and vanilla SAM on the previously mentioned four tasks. According to the figure, our Med-SA model consistently produces better segmentation masks than the original SAM across all four modalities on all prompt settings. The original SAM only performs well on the simplest melanoma segmentation task with more accurate prompts like 3-points and BBox 0.75. On the contrary, Med-SA performs well in most cases, only failing to segment accurately on optic disc and cup segmentation with a BBox 0.5 prompt. This again proves the need for fine-tuning SAM for medical images and the generalization ability of Med-SA over various segmentation tasks and image modalities.

4.6. Effectiveness of key components

We conduct a comprehensive ablation study to validate the effectiveness of the proposed SD-Trans and HyP-Adpt. The results are presented in Table 5, where the baseline (first line) represents a simple combination of SAM and the original Adaption method. In the baseline setting, 3D images are treated as a sequence of 2D images and processed individually, without involving prompts in the Adaption process. As shown in the table, our 2D to 3D design significantly enhances the performance compared to the vanilla SAM plus Adaption setting on both 3D data benchmarks (BTCV and BrainTumor). This improvement highlights the effectiveness of our proposed 2D to 3D design. In the Prompt-conditional Adaption, we compare HyP-Adpt with two simpler alternatives: addition and concatenation, for combining the prompt embedding. While addition and concatenation also show some effectiveness, the improvements achieved are still marginal. On the other hand, using the proposed HyP-Adpt leads to a significant enhancement in performance, further validating the effectiveness of our proposed HyP-Adpt design.

4.7. Detailed analysis

Different Positions of adapter. We commence by analysing specific positions to introduce the Adapter, and the results are summarized in Table 6. We compare the introduction of adapters on the main branch, the MLP residue branch, and a combination of both. Our findings indicate that introducing adapters on either the main branch or the MLP residual branch yields comparable results. However, introducing adapters on both branches achieves improved performance, demonstrating the effectiveness of inserting adapters on both the main branch and MLP residual path.

Different Hyper-parameters in adapter. We do comprehensive experiments on adapter ranks, numbers, and scaling factors to find the best hyper-parameters setting ensuring the best performance of Med-SA. Concretely, we vary the number of ranks (middle dimension) of the adapters, and scaling factor, and explore the insertion of adapters on different numbers of layers. The results on the BTCV dataset are presented in Table 7. In sub-table (a), we observe a consistent improvement in accuracy as the middle dimension increases up to 64, reaching a saturation point at approximately 64. Notably, Med-SA demonstrates decent performance even with a reduced number of

Table 8
Comparison of information fusion methods in the SD-Trans module.

Method	#Params	BTCV		Optic-Disc	
		Dice (%)	IoU (%)	Dice (%)	IoU (%)
Add.	0	87.8	79.8	96.8	88.8
Concat.	0	87.4	79.1	96.1	88.3
Cross-Atten.	28.3M	**88.3**	**80.6**	**97.4**	**89.5**

ranks, achieving a Dice Score of about 75.2% with only one rank. Sub-table (b) demonstrates a positive correlation between the number of adapted layers and Med-SA performance. Notably, when adding the same number of layers, Med-SA exhibits a preference for the top part of the network over the bottom part. For instance, Med-SA with 17–32 achieves approximately 9% higher Dice than 1–16, both equipped with 16 adapted layers. To balance the features generated by the original frozen branch and the tunable bottleneck branch, we introduce the scaling factor s. As shown in sub-table (c), evaluating Med-SA with multiple s values on both 2D and 3D segmentation tasks reveals the consistent optimal results when $s = 0.05$, deviating from preferences observed in NLP (larger s) (He et al., 2021), but aligns with the result in natural images (Chen et al., 2022; Yang et al., 2024). This empirical choice may be attributed to the training stability, as larger scaling factors in the vision domain may cause the network to become overly reliant on the adapter during training. SAM is trained with massive data and has strong segmentation ability, and the training process may be unstable with larger scaling factors. Therefore, even a small scaling factor enables SAM to effectively perform medical segmentation tasks, and we choose $s = 0.05$ as a default setting.

Different Design of SD-Trans Module To further investigate how to combine information from the spatial and depth branches in the SD-Trans module, we compared three methods for merging the outputs of the two branches: adding, concatenation, and cross-attention. As shown in Table 8, direct adding outperforms concatenation, while cross-attention achieves the best results. However, employing cross-attention introduces 28.3M additional parameters, and is even more then the origin adapter's parameter (13.1M). This may reduce inference speed and conflicts with the lightweight design philosophy of the adapter. Given that adding provides sufficiently effective information fusion, we adopt adding as the fusion method in the SD-Trans module.

Generic Nature of Med-SA Framework. The proposed Med-SA is a generic framework that can easily accept many SAM variants. We verify the method with several SAM variants, especially the lightweight ones which are crucial for medical applications. Namely, MobileSAM (Zhang et al., 2023a) and EfficientSAM (Xiong et al., 2023), which have emerged as solutions. To assess their efficiency and performance, we conduct ablation experiments by applying our proposed methods to two scales of the origin SAM (ViT-H and ViT-B) and these lightweight SAM versions. The results on optic-disc and melanoma segmentation tasks are presented in Table 9. It can be observed from the table that these lightweight SAM variants seamlessly integrate into our framework and achieve comparable, or even superior, performance with reduced GPU memory consumption. It is worth noting that although SAM (ViT-H) achieves the best performance overall, MobileSAM outperforms SAM (ViT-B) by 2.45% and 3.89% on Dice and IoU scores on OpticeDisc task, indicating our framework's potential for different backbones.

Inference Cost of Med-SA The proposed Med-SA is a highly efficient framework that updates only an extra 2% of the parameters in the original SAM. To evaluate the impact of these additional parameters, we compare the inference speed of Med-SA with the original SAM on both 2D and 3D segmentation tasks, using an A6000 GPU for testing. As shown in Fig. 5, Med-SA significantly enhances the performance of 2D and 3D medical image segmentation, with only a slight increase in inference time. Notably, Med-SA achieves the best performance when

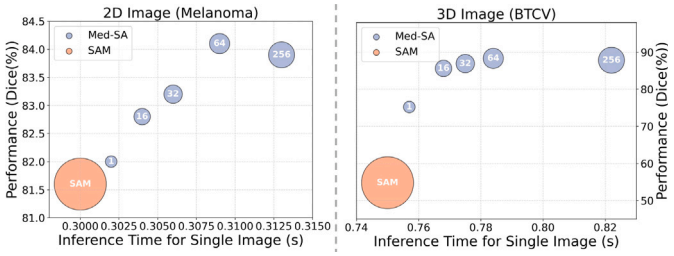


Fig. 5. The accuracy and inference time of different adapter ranks. The white numbers within the circles indicate the adapter ranks. The scale of the circles indicates the scale of parameters.

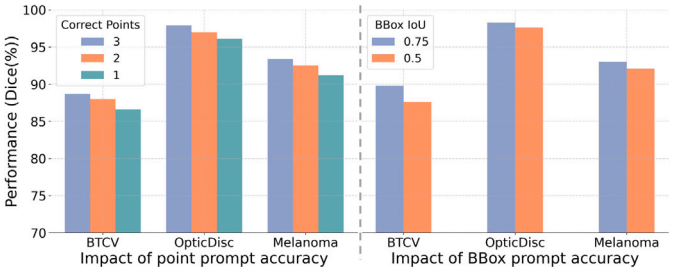


Fig. 6. The impact of prompt accuracy on segmentation performance.

the adapter rank is set to 64, surpassing the original SAM by 2.5% and 33.5% on the BTCV and Melanoma datasets, respectively, while the inference time increases by only 0.008 s and 0.034 s. Furthermore, increasing the adapter rank has minimal impact on inference speed, as the additional parameters are negligible compared to the overall size of the original SAM.

5. Discussion

5.1. The algorithm's robustness

In the experiment of interactive models, the use of random points and bounding boxes inevitably brings randomness. To further evaluate the impact of random prompts, we repeated the experiments five times on all medical tasks, and present the results of three tasks in Table 10. As shown in the table, the variances across all datasets are small. This indicates that when datasets contain a large number of examples, the impact of random prompts on the final results is minimal, demonstrating the robustness of our Med-SA.

Additionally, we further investigate the impact of incorrect prompts on the experimental results. Our experiment setting includes both the point prompt and the bounding box prompt. For point prompts, we analysed scenarios with 1, 2, and 3 correct points (located within the target area) when using 3 points in total as input. For bounding box prompts, we considered the original settings of BBox 0.5 and BBox 0.75 since these settings also provide insights into the impact of prompt accuracy. As shown in Fig. 6, as prompt accuracy decreases, the segmentation performance of the model shows a slight decline but still maintains a high overall performance. This result is somehow surprising, considering that the original SAM is a prompt-based method and the performance of fully finetuned methods drops rapidly when the prompt accuracy declines. As shown in Table 2, when the accuracy of the bounding box prompt decreases (BBox 0.75 to BBox 0.5), the average Dice scores of the original SAM, SAMed, SAM3D, and MedSAM drop by 11.0%, 14.4%, 4.2%, and 11.2% on the BTCV dataset, respectively. We attribute this improvement to our proposed task-specific adapter, which effectively identifies specific semantic regions. This enables the model to accurately recognize the target organ or tissue, even when the prompts are not entirely precise, thereby reducing the dependency on human prompts.

Table 9
Different SAM variants applied in Med-SA evaluated by Dice score (%) and mIoU (%). Best results are denoted in **bold**.

Backbones	Memory	Optic-Disc		Optic-Cup		Melanoma	
		Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)
SAM (ViT-H)	29462M	**97.4**	**89.5**	86.8	78.8	**92.6**	84.1
SAM (ViT-B)	16368M	90.9	84.2	80.2	68.5	92.3	**86.5**
EfficientSAM (ViT-S)	15662M	80.0	69.5	74.3	62.4	91.0	84.4
MobileSAM (Tiny-ViT)	10178M	93.3	88.1	82.9	73.0	92.3	**86.5**

Table 10
Standard deviation across five repeated experiments for all settings of Med-SA.

Method	BTCV		OpticDisc		Melanoma	
	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
Med-SA everything	0.8	0.8	0.7	0.5	0.7	0.6
Med-SA 1 point	0.6	0.5	0.5	0.4	0.5	0.5
Med-SA 3 points	0.4	0.3	0.2	0.2	0.3	0.2
Med-SA BBox 0.5	0.8	0.7	0.6	0.5	0.8	0.7
Med-SA BBox 0.75	0.3	0.3	0.2	0.1	0.2	0.1

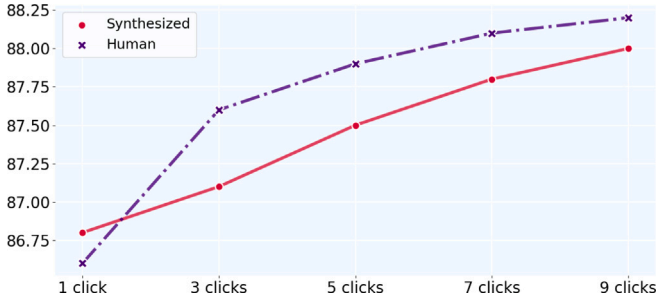


Fig. 7. Comparison of clinician and synthesized prompts on optic-cup segmentation task, evaluated by Dice.

5.2. Human prompt v.s. Synthesized prompt

To test how well Med-SA performs in real-world clinics, we have certified clinicians use it and compare their prompts with synthesized prompts for optic-cup segmentation in fundus images. Three senior ophthalmologists, each with over a decade of experience, provided the prompts. The comparison is shown in Fig. 7, which shows that while synthesized prompts were more accurate with a single click, human-generated prompts became superior with additional clicks. This observation aligns with the tendency of humans to initially click on the centre of the object, providing limited information on its segmentation boundaries. However, when prompted for more input, humans tend to click on the ambiguous edges of the object, offering more detailed information than random synthesized clicks. This also suggests clinicians may use three clicks for the optimal balance in practical use.

5.3. Failure case and future work

While Med-SA has already demonstrated promising results, there remains much to explore and improve, particularly in processing ambiguous prompts. For instance, in certain scenarios, users might unavoidably provide ambiguous prompts that do not distinctly identify target objects, as illustrated in Fig. 8. In this particular image, due to the pancreas's complex structure, click prompts demand considerable effort. However, opting for a bounding box to encircle the pancreas inevitably results in also encompassing the aorta. In such instances, Med-SA segments all potential organs, allowing users to make a selection in the subsequent step, which could be cumbersome. To address this issue, future enhancements could involve adapting the network to accept a broader array of prompt types, such as scribbles and text, enabling users to specify their targets more easily in complex situations.

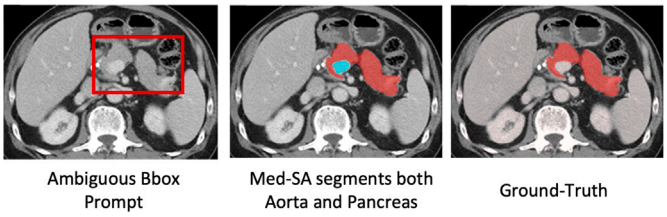


Fig. 8. A case that the users have to contain aorta while intending to box the pancreas. Med-SA would fail under ambiguous prompts by segmenting all possible organs.

Table 11
Comparison with MedSAM-2 evaluated by Dice score (%).

Method	Prompt type	Prms (M)	Tunable Prms (M)	BTCV	Optic-Cup
Med-SA	1 point	636	13	88.3	86.8
	3 point	636	13	88.7	87.1
	BBox 0.5	636	13	87.6	86.4
	BBox 0.75	636	13	89.8	87.5
MedSAM-2	Single prompt	224.4	224.4	89.0	79.9

Recently, SAM2 (Ravi et al., 2024) was proposed as a strong backbone for video segmentation. MedSAM-2 (Zhu et al., 2024) adapts SAM2 to the medical domain through full fine-tuning, treating a 3D image as a video sequence and providing a single prompt for the first slice, thus requiring fewer prompts. For 2D images, it similarly treats the entire dataset as a video clip and provides a prompt only for the first image. We have also compared the parameters, prompt type, and performance between Med-SA and MedSAM-2 (trained on 64 medical datasets and conduct zero-shot testing), the results are shown below:

As shown in Table 11, Med-SA still outperforms MedSAM-2, but MedSAM-2 also achieves competitive results. Notably, MedSAM-2 demonstrates strong performance on the 3D BTCV dataset, with its average Dice score surpassing all prompt settings of Med-SA, except for BBox 0.75. Considering that MedSAM-2 requires significantly fewer prompts than Med-SA, it holds great potential in the medical domain. However, fully fine-tuning MedSAM-2 still requires updating a large number of parameters. In future work, we plan to apply adapter techniques to MedSAM-2 to improve its efficiency.

6. Conclusion

In this paper, we extend SAM, a powerful general segmentation model, to address medical image segmentation, introducing Med-SA. Using parameter-efficient adaptation with simple yet effective SD-Trans and HyP-Adpt, we achieve substantial improvements over the original

SAM model. Our approach result in SOTA performance across 17 medical image segmentation tasks spanning 5 different image modalities. We anticipate this work will serve as a stepping stone towards advancing foundation medical image segmentation and inspire the development of novel fine-tuning techniques.

CRediT authorship contribution statement

Junde Wu: Writing – original draft, Project administration, Methodology, Formal analysis, Data curation. **Ziyue Wang:** Writing – review & editing, Validation, Software. **Mingxuan Hong:** Writing – review & editing, Conceptualization. **Wei Ji:** Writing – review & editing, Software, Methodology. **Huazhu Fu:** Writing – review & editing, Methodology. **Yanwu Xu:** Validation. **Min Xu:** Writing – review & editing, Methodology. **Yueming Jin:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgements

This work was supported by Ministry of Education Tier 1 Start up grant, NUS, Singapore (A-8001267-01-00); Ministry of Education Tier 1 grant, NUS, Singapore (A-8001946-00-00).

Data availability

All datasets are publicly available. We will release the code, and the code link is included in the paper.

References

Amit, T., Nachmani, E., Shaharbany, T., Wolf, L., 2021. Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390.

Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al., 2021. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314.

Bui, N.-T., Hoang, D.-H., Tran, M.-T., Le, N., 2023. Sam3d: Segment anything model in volumetric medical images. arXiv preprint arXiv:2309.03493.

Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P., 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. arXiv preprint arXiv:2205.13535.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021a. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al., 2024. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med. Image Anal. 97, 103280.

Chen, C., Miao, J., Wu, D., Yan, Z., Kim, S., Hu, J., Zhong, A., Liu, Z., Sun, L., Li, X., et al., 2023. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. arXiv preprint arXiv:2309.08842.

Chen, X., Zhao, Z., Yu, F., Zhang, Y., Duan, M., 2021b. Conditional diffusion for interactive segmentation. In: ICCV. pp. 7345–7354.

Cheng, Z., Wei, Q., Zhu, H., Wang, Y., Qu, L., Shao, W., Zhou, Y., 2024. Unleashing the potential of SAM for medical adaptation via hierarchical decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3511–3522.

Cheng, J., et al., 2023. Sam-med2d. arXiv preprint arXiv:2308.16184.

Chu, C., Zheng, J., Zhou, Y., 2021. Ultrasonic thyroid nodule detection method based on U-Net network. Comput. Methods Programs Biomed. 199, 105906.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. Springer, pp. 424–432.

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., et al., 2023b. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. arXiv preprint arXiv:2304.04155.

Deng, X., Wu, H., Zeng, R., Qin, J., 2024. MemSAM: Taming segment anything model for echocardiography video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9622–9631.

Deng, G., et al., 2023a. SAM-U: Multi-box prompts triggered uncertainty estimation for reliable SAM in medical image. arXiv preprint arXiv:2307.04973.

Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., 2023. QLoRA: Efficient finetuning of quantized LLMs. arXiv preprint arXiv:2305.14314.

Dong, Y., Cordonnier, J.-B., Loukas, A., 2021. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In: International Conference on Machine Learning. PMLR, pp. 2793–2803.

Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fang, H., Li, F., Fu, H., Sun, X., Cao, X., Son, J., Yu, S., Zhang, M., Yuan, C., Bian, C., et al., 2022. REFUGE2 challenge: Treasure for multi-domain learning in glaucoma assessment. arXiv preprint arXiv:2202.08994.

Fang, X., Yan, P., 2020. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Trans. Med. Imaging 39 (11), 3619–3629.

Gong, H., Chen, G., Wang, R., Xie, X., Mao, M., Yu, Y., Chen, F., Li, G., 2021. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 257–261.

Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.-A., Dou, Q., 2023. 3Dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. arXiv preprint arXiv:2306.13465.

Grady, L., 2006. Random walks for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 28 (11), 1768–1783.

Gu, A., Dao, T., 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Gulshan, V., Rother, C., Criminisi, A., Blake, A., Zisserman, A., 2010. Geodesic star convexity for interactive image segmentation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, pp. 3129–3136.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2022a. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. Springer, pp. 272–284.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022b. Unetr: Transformers for 3d medical image segmentation. In: WACV. pp. 574–584.

He, X., Li, C., Zhang, P., Yang, J., Wang, X.E., 2022. Parameter-efficient fine-tuning for vision transformers. arXiv preprint arXiv:2203.16329.

He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G., 2021. Towards a unified view of parameter-efficient transfer learning. arXiv preprint arXiv:2110.04366.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Adv. Neural Inf. Process. Syst. 33, 6840–6851.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al., 2024. Segment anything model for medical images? Med. Image Anal. 92, 103061.

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

Jie, S., Deng, Z.-H., 2023. Fact: Factor-tuning for lightweight adaptation on vision transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1060–1068.

Kim, B., Oh, Y., Ye, J.C., 2022. Diffusion adversarial representation learning for self-supervised vessel segmentation. arXiv preprint arXiv:2209.14566.

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

Li, Z., Chen, Q., Koltun, V., 2018. Interactive image segmentation with latent diversity. In: CVPR. pp. 577–585.

Li, H., Liu, H., Hu, D., Wang, J., Oguz, I., 2023. Promise: Prompt-driven 3d medical image segmentation using pretrained image foundation models. arXiv preprint arXiv:2310.19721.

Lin, Z., Zhang, Z., Chen, L.-Z., Cheng, M.-M., Lu, S.-P., 2020. Interactive image segmentation with first click attention. In: CVPR. pp. 13339–13348.

Liu, Y., Lu, Z., Li, J., Yang, T., Yao, C., 2019. Deep image-to-video adaptation and fusion networks for action recognition. IEEE Trans. Image Process. 29, 3168–3182.

Liu, J., Yang, H., Zhou, H.-Y., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., et al., 2024. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 615–625.

Ma, J., Li, F., Wang, B., 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722.

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

Ma, J., Wu, F., Jiang, T., Zhao, Q., Kong, D., 2017. Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. Int. J. Comput. Assist. Radiol. Surg. 12 (11), 1895–1910.

Mahadevan, S., Voigtlaender, P., Leibe, B., 2018. Iteratively trained interactive segmentation. arXiv preprint arXiv:1805.04398.

Milton, M.A.A., 2019. Automated skin lesion classification using ensemble of deep neural networks in isic 2018: Skin lesion analysis towards melanoma detection challenge. arXiv preprint arXiv:1901.10802.

Paranjape, J.N., Nair, N.G., Sikder, S., Vedula, S.S., Patel, V.M., 2024. Adaptivesam: Towards efficient tuning of sam for surgical scene segmentation. In: Annual Conference on Medical Image Understanding and Analysis. Springer, pp. 187–201.

Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E., 2015. An open access thyroid ultrasound image database. In: 10th International Symposium on Medical Information Processing and Analysis. Vol. 9287, International Society for Optics and Photonics, 92870W.

Qiu, J., Li, L., Sun, J., Peng, J., Shi, P., Zhang, R., Dong, Y., Lam, K., Lo, F.P.-W., Xiao, B., et al., 2023. Large ai models in health informatics: Applications, challenges, and the future. IEEE J. Biomed. Heal. Inform..

Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging. Adv. Neural Inf. Process. Syst. 32.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al., 2024. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. Springer, pp. 234–241.

Roy, S., Wald, T., Koehler, G., Rokuss, M.R., Disch, N., Holzschuh, J., Zimmerer, D., Maier-Hein, K.H., 2023. SAM. MD: Zero-shot medical image segmentation capabilities of the segment anything model. arXiv preprint arXiv:2304.05396.

Song, J., Meng, C., Ermon, S., 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502.

Song, Y., Zhou, Q., Li, X., Fan, D.-P., Lu, X., Ma, L., 2024. BA-SAM: Scalable bias-mode attention mask for segment anything model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3162–3173.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J., 2021b. Transbts: Multimodal brain tumor segmentation using transformer. In: MICCAI. Springer, pp. 109–119.

Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y., 2023. SAM-Med3D. arXiv:2310.15161.

Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., Qin, J., 2021a. Boundary-aware transformers for skin lesion segmentation. In: MICCAI. Springer, pp. 206–216.

Wang, S., Yu, L., Li, K., Yang, X., Fu, C.-W., Heng, P.-A., 2019. Boundary and entropy-driven adversarial learning for fundus image segmentation. In: MICCAI. Springer, pp. 102–110.

Wei, Q., et al., 2023. Focused and collaborative feedback integration for interactive image segmentation. In: CVPR. pp. 18643–18652.

Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C., 2021. Diffusion models for implicit image segmentation ensembles. arXiv preprint arXiv:2112.03145.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S., 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: CVPR. pp. 16133–16142.

Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., Wen, Z., 2022a. FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. Med. Image Anal. 76, 102327.

Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y., 2022b. MedSegDiff: Medical image segmentation with diffusion probabilistic model. arXiv preprint arXiv:2211.00611.

Xie, Y., Richmond, D., 2018. Pre-training on grayscale imagenet improves medical image classification. In: ECCV Workshops.

Xie, Y., Zhang, J., Shen, C., Xia, Y., 2021. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. Springer, pp. 171–180.

Xing, Z., Wan, L., Fu, H., Yang, G., Zhu, L., 2023. Diff-UNet: A diffusion embedded network for volumetric segmentation. arXiv preprint arXiv:2303.10326.

Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., Krishnamoorthi, R., Chandra, V., 2023. EfficientSAM: Leveraged masked image pretraining for efficient segment anything. arXiv:2312.00863.

Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.S., 2016. Deep interactive object selection. In: CVPR. pp. 373–381.

Yan, X., Sun, S., Han, K., Le, T.-T., Ma, H., You, C., Xie, X., 2024. AFTer-SAM: Adapting SAM with axial fusion transformer for medical imaging segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7975–7984.

Yang, L., Zhang, R.-Y., Wang, Y., Xie, X., 2024. MMA: Multi-modal adapter for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23826–23837.

Yeh, S.-Y., Hsieh, Y.-G., Gao, Z., Yang, B.B., Oh, G., Gong, Y., 2023. Navigating text-to-image customization: From lycoris fine-tuning to model evaluation. In: The Twelfth International Conference on Learning Representations.

Yu, S., Xiao, D., Frost, S., Kanagasingam, Y., 2019. Robust optic disc and cup segmentation with deep learning for glaucoma detection. Comput. Med. Imaging Graph. 74, 61–71.

Yue, W., Zhang, J., Hu, K., Xia, Y., Luo, J., Wang, Z., 2024. Surgicalsam: Efficient class promptable surgical instrument segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 6890–6898.

Zaken, E.B., Ravfogel, S., Goldberg, Y., 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:2106.10199.

Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., Zhao, T., 2023c. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv:2303.10512.

Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S.-H., Lee, S., Hong, C.S., 2023a. Faster segment anything: Towards lightweight SAM for mobile applications. arXiv preprint arXiv:2306.14289.

Zhang, S., Metaxas, D., 2023. On the challenges and perspectives of foundation models for medical image analysis. Med. Image Anal. 102996.

Zhang, Y., Shen, Z., Jiao, R., 2024. Segment anything model for medical image segmentation: Current applications and future directions. Comput. Biol. Med. 108238.

Zhang, K., et al., 2023b. Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785.

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer, pp. 3–11.

Zhou, T., et al., 2023. Volumetric memory network for interactive medical image segmentation. Med. Image Anal. 83, 102599.

Zhu, J., Qi, Y., Wu, J., 2024. Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874.