RAG-SAM: A Hybrid Framework for Medical Image Segmentation Using Retrieval-Augmented Generation and Segment Anything Model

Yuanjie Chen, Zichen Zou, Yuan Gao, Huayi Wang
Shanghai Jiao Tong University
{cyj2003, zzcnb123456, yoggy23, why618188}@sjtu.edu.cn

Abstract

Medical image segmentation is essential for diagnosis and treatment planning but faces challenges due to limited annotated data and domain-specific requirements. To address these issues, we explore parameter-efficient fine-tuning techniques, Low-Rank Adaptation(LoRA) and Adapter tuning, to adapt the Segment Anything Model (SAM) for medical imaging tasks with minimal parameter updates. We conduct experiments on the BTCV and FLARE22 datasets, demonstrating the effectiveness of these methods in improving segmentation accuracy while maintaining computational efficiency. To leverage few-shot capability of SAM, we integrate the Retrieval-Augmented Generation(RAG) framework. Our work provides an efficient approach to adapt foundation models like SAM for medical image segmentation, achieving state-of-the-art performance with significantly reduced computational overhead.

1. Introduction

Medical image segmentation is essential for tasks like diagnosis and treatment planning, enabling clinicians to identify anatomical structures and abnormalities. However, it faces challenges such as limited data and annotations, as generating accurate labels requires trained medical professionals and is time-consuming. This results in a small number of labeled samples, making it difficult to train models effectively. The primary challenge is utilizing this limited annotated data to produce accurate segmentation masks without compromising performance.

Foundation models, trained on large-scale datasets from diverse domains, have shown impressive performance in few-shot and zero-shot scenarios. In the field of natural image segmentation, Segment Anything Model(SAM)[11] has shown remarkable zero-shot performance in accuracy and strong generalization ability, suggesting its potential for segmentation in medical images. Though SAM performs well on natural image segmentation tasks, its performance in the medical image domain is not satisfactory.

To address the challenges of limited annotations and domain-specific requirements in medical image segmentation, we explored two parameter-efficient fine-tuning (PEFT) techniques: Low-Rank Adaptation[21] (LoRA) and Adapter tuning[20]. These techniques are designed to adapt large foundation models, such as SAM, to specific downstream tasks with minimal parameter updates, making them well-suited for resource-constrained scenarios like medical imaging. LoRA reduces the number of trainable parameters by introducing low-rank decompositions to specific weight matrices, enabling efficient fine-tuning with limited computational cost. Similarly, Adapter tuning achieves adaptation by inserting lightweight modules into the pre-trained model, allowing task-specific customization while keeping the majority of the model frozen.

We conduct comprehensive experiments on two prominent medical image segmentation benchmarks. First, we compared the performance of LoRA and Adapter to other medical segmentation Models on the widely-used BTCV dataset [12], which focuses on multi-organ segmentation in abdominal CT scans. This comparison aimed to determine the relative strengths and weaknesses of these two parameter-efficient fine-tuning techniques when adapting SAM to medical imaging tasks. Second, we performed a detailed parameter ablation study about rank size of LoRA matrices on the FLARE22 dataset[16]. This ablation experiment analyzed the impact of varying LoRA's rank size on segmentation performance, offering insights into the trade-off between model efficiency and accuracy. Together, these experiments provide a robust evaluation of the feasibility and adaptability of PEFT techniques for medical image segmentation.

To further expand SAM's applicability in the medical domain, we deploy the Retrieval-Augmented Generation (RAG)[15] framework to our fine-tuned medical-specific SAM. This framework enables the fine-tuned SAM model to effectively utilize its few-shot capabilities by incorporating relevant contextual information retrieved from external sources, thereby improving its performance in medical image segmentation tasks.

Our work makes the following contributions to the field of medical image segmentation:

- We explore two parameter-efficient fine-tuning techniques, **LoRA** and **Adapter**, to adapt SAM for medical image segmentation with minimal parameter updates.
- We conduct experiments on the **BTCV** and **FLARE22** , comparing LoRA and Adapter performance and analyzing LoRA's parameter efficiency through ablation studies.
- We integrate the **Retrieval-Augmented Generation (RAG)** into SAM, activating its few-shot capabilities for improved generalization to unseen medical tasks.
- Our work demonstrates an efficient and scalable approach to adapt foundation models like SAM for medical image segmentation.

2. Related Work

2.1. Medical Image Segmentation Models

Medical image segmentation is a fundamental task in medical image analysis, playing a critical role in applications such as diagnosis, treatment planning, and monitoring. Over the years, numerous deep learning-based models have been proposed to tackle this challenge, leveraging advancements in neural network architectures and attention mechanisms.

One of the pioneering architectures in this domain is U-Net, which utilizes a symmetric encoder-decoder structure with skip connections to capture both global context and fine-grained details. Building upon this foundation, several advanced models have been introduced to further enhance segmentation accuracy and robustness.

UNETR (U-Net Transformer) [7] integrates the Transformer architecture into the U-Net framework to capture long-range dependencies in volumetric medical images. By utilizing a self-attention mechanism in the encoder, UNETR has demonstrated state-of-the-art performance on tasks such as brain and liver segmentation.

EnsDiff [19] employs a diffusion-based framework combined with ensemble learning to improve the stability and accuracy of medical image segmentation, particularly in noisy and challenging datasets. The model generates diverse segmentation predictions and combines them to achieve more robust results.

TransUNet [4] merges the strengths of Transformers and convolutional neural networks (CNNs) by embedding Transformer blocks into the U-Net structure. This hybrid architecture effectively captures both local features via CNNs and global dependencies via the Transformer, achieving impressive results in organ segmentation tasks.

2.2. Foundation Models

Foundation models have emerged as a transformative paradigm in machine learning, enabling the development of versatile, general-purpose systems that can be fine-tuned for a wide range of downstream tasks. In the domain of image segmentation, the Segment Anything Model (SAM) has become a prominent example of this approach.

Segment Anything Model (SAM) is a foundational model designed to perform zero-shot image segmentation across diverse datasets and tasks. SAM leverages a powerful combination of a vision transformer (ViT) backbone and a pre-trained promptable interface, which allows users to specify segmentation tasks via points, bounding boxes, or text-based prompts. The model's generalization capability stems from its extensive training on a large-scale dataset containing over 1 billion masks, covering a wide range of image domains.

SAM has demonstrated remarkable performance in medical image segmentation tasks, despite being developed as a general-purpose segmentation tool. By leveraging some strategies like fine-tuning and RAG, we can make foundation model SAM more powerful in medical image segmentation.

2.3. Finetuning Strategies

To adapt the Segment Anything Model (SAM) for medical image segmentation, we leverage two fine-tuning strategies: Low-Rank Adaptation (LoRA) and Adapter-based fine-tuning. These parameter-efficient techniques enable effective customization of SAM for domain-specific tasks while minimizing computational and storage overhead.

Low-Rank Adaptation (LoRA) focuses on updating a small subset of SAM's parameters by introducing low-rank decompositions in the model's transformer blocks. Specifically, LoRA adds lightweight trainable layers to the frozen parameters of the pre-trained SAM, allowing efficient optimization without compromising the model's original capabilities. By fine-tuning SAM's image encoder with LoRA, this approach enables the extraction of medical-specific features, improving segmentation accuracy. In the SAMed framework, LoRA is applied to the query and value projection layers of the transformer blocks, achieving high segmentation performance with a significantly reduced computational footprint.

Adapter-based Fine-tuning is another parameter-efficient technique employed to adapt SAM for medical applications. The Medical SAM Adapter (Med-SA) introduces lightweight adapter modules into the frozen layers of SAM, specifically within the image encoder and mask decoder. These adapters consist of down-projection and up-projection layers, allowing domain-specific fine-tuning with minimal parameter updates. Additionally, Med-SA incorporates techniques such as Space-Depth Transpose (SD-Trans) to handle 3D medical imaging modalities and Hyper-Prompting Adapter (HyP-Adpt) for prompt-conditioned adaptation. This approach demonstrates strong

performance across various medical image segmentation tasks, outperforming fully fine-tuned counterparts while only updating 2% of the parameters.

Both LoRA and Adapter-based fine-tuning showcase the effectiveness of parameter-efficient strategies in adapting large-scale foundation models like SAM for medical image segmentation. These methods not only reduce the computational and storage costs but also retain the generalization capabilities of the original model, making them highly suitable for clinical applications.

2.4. Retrieval-Augmented Techniques

Retrieval-Augmented Generation (RAG) in natural language processing integrates the advantages of retrieval-based systems and generative models, leveraging the strengths of both approaches to enhance performance and accuracy [13], [6]. RAG operates by first retrieving relevant information from an external knowledge source based on the input query. This retrieved information provides contextual grounding, enhancing the accuracy, informativeness, and relevance of the responses generated by large language models (LLMs). In the image domain, retrieval-augmented techniques have been applied to various tasks. For instance, the Retrieval-Augmented Diffusion Model (RDM) has been introduced for image synthesis, where the generative model is conditioned on relevant samples retrieved from a database to enhance the quality and coherence of the generated images [2]. Similarly, Retrieval-Augmented Classification (RAC) incorporates an explicit retrieval module into standard image classification pipelines, enhancing performance, particularly in long-tail visual recognition tasks [15]. Additionally, Retrieval-Augmented Customization (REACT) utilizes relevant image-text pairs from a large-scale web database to tailor visual models for specific target domains, achieving notable performance improvements across tasks while reducing the need for extensive retraining [14]. In this work, we extend the retrieval-based approach to the image segmentation domain by designing a retrieval module that queries an external database for contextual and anatomical information. This retrieved information is then used to guide the SAM2 model in few-shot medical image segmentation, eliminating the need for any retraining.

er Times is specified, Times Roman may also be used.

3. Method

3.1. Fine-tune with LoRA

We first fine-tune the large-scale image segmentation model, Segment Anything Model (SAM), with the LoRA method [9]. This approach is elaborated on **SAMed** [21]. In the context of SAM, we adopt this strategy for efficient and stable fine-tuning of specific model components.

The model architecture of SAMed, as shown in Figure 1,

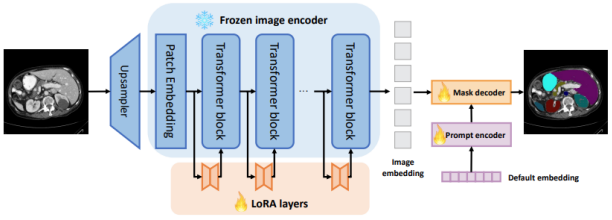

Figure 1. Overview of SAMed architecture with LoRA integration. The model adopts the structure of SAM and incorporates low-rank approximations via LoRA to enable efficient fine-tuning of the transformer layers.

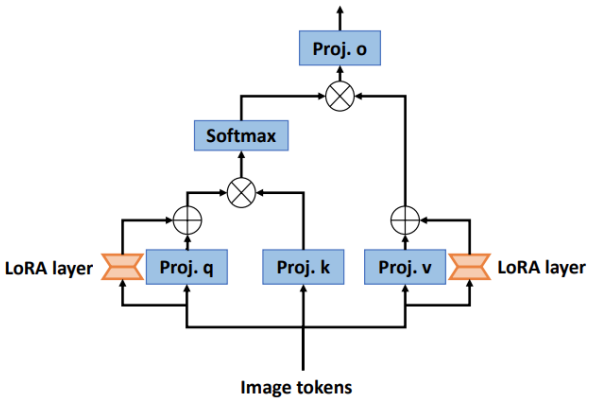

Figure 2. Illustration of the LoRA strategy in SAMed. The low-rank update for the weight matrix W is achieved through an additional bypass, consisting of two linear layers A and B. The dimensions of these layers are chosen such that $r \ll \min\{C_{in}, C_{out}\}$.

inherits from SAM, where we freeze all parameters of the image encoder. We then design a trainable bypass for each transformer block. In accordance with LoRA, these bypasses first condense the transformer features into a low-rank space and then reproject the reduced features to align with the output feature channels of the frozen transformer blocks. The prompt encoder in SAMed does not require any additional prompts during inference for automatic segmentation, which greatly benefits its application in automatic medical diagnosis.

The mask decoder in SAM consists of a lightweight transformer decoder and a segmentation head. Fine-tuning the transformer decoder with LoRA is optional in SAMed, and we explore this option to determine its impact .

More specifically, the LoRA strategy in SAMed is illustrated in Figure 2. Given the encoded token sequence $F \in \mathbb{R}^{B\times N\times C_{in}}$ and the output token sequence $\hat{F} \in \mathbb{R}^{B\times N\times C_{out}}$ processed by a projection layer $W \in \mathbb{R}^{C_{out}\times C_{in}}$, LoRA assumes the weight update for W should be gradual and stable. To achieve this, a low-rank approximation is used. SAMed first freezes the transformer layers,

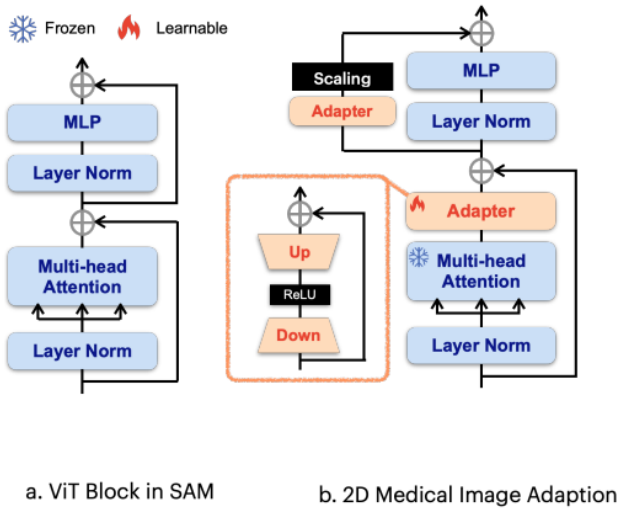

Figure 3. Illustration of the Medical SAM Adapter (MSA). The figure shows the insertion of Adapter modules into the ViT blocks of SAM for the medical imaging task.

keeping W fixed, and then adds a bypass to perform the low-rank approximation. This bypass consists of two linear layers, $A \in \mathbb{R}^{r \times C_{in}}$ and $B \in \mathbb{R}^{C_{out} \times r}$, where r is a small value relative to C_{in} and C_{out}. Thus, the processing of the updated weight matrix $\hat{W}$ can be described as:

$$\hat{F} = \hat{W}F, \tag{1}$$
$$\hat{W} = W + \Delta W = W + BA$$

We observe that SAMed achieves improved performance when LoRA is applied to the query and value projection layers in the multi-head self-attention mechanism.

Finally, the mask decoder in SAM is composed of a lightweight transformer layer and a segmentation head. It is optional to apply LoRA to the transformer layer and fine-tune the segmentation head, or to fine-tune all parameters in the mask decoder directly. Both strategies are viable in terms of training and deployment overhead. The second approach, while resulting in a smaller model size, may lead to lower performance, especially for tasks requiring high accuracy. The mask decoder of SAMed predicts k semantic masks $\hat{S}_l \in \mathbb{R}^{h \times w \times k}$, each corresponding to a different semantic label. The final segmentation map is produced as:

$$\hat{S} = \arg\max(\mathrm{Softmax}(\hat{S}_l), d = -1), \tag{2}$$

where $d = -1$ indicates that the Softmax and Argmax operations are performed along the last dimension, which corresponds to the channel dimension.

3.2. Fine-tune with Adaption

In addition to LoRA, we explore another fine-tuning strategy for SAM, known as Adaption [9], which is particularly effective for domain-specific tasks such as medical image segmentation. This approach is elaborated on **Med-SA** [20]. The core idea behind Adaption is to insert Adapter modules into the pre-trained model and only update a small set of additional parameters while keeping the majority of the model frozen. These modules enable the model to adapt to new tasks without the need for fine-tuning all of its parameters.

In Med-SA, we utilize two Adapter modules for each Vision Transformer (ViT) block in the encoder. For a standard ViT block, as depicted in Figure 3(a), the first Adapter is inserted after the multi-head attention and before the residual connection, as shown in Figure 3(b). The second Adapter is placed in the residual path of the multi-layer perceptron (MLP) layer, following the multi-head attention. To improve the performance further, we apply a scaling factor s to the output of the second Adapter, as proposed by [5].

In the SAM decoder, we follow a similar strategy, where we fine-tune all parameters in the mask decoder using the Adaption method, as done with the LoRA strategy. However, depending on the specific use case, it is possible to fine-tune only the adapter parameters while keeping the rest of the decoder frozen.

3.3. Few-shot Segmentation Framework

Our few-shot segmentation framework, illustrated in Figure 4, is specifically designed to address medical image segmentation tasks with limited annotated data by incorporating a retrieval-augmented system. Built upon the Segment Anything Model 2 (SAM 2) [18], this framework operates without the need for additional training, ensuring high adaptability and efficiency across a range of medical imaging applications. Given an input image, the framework begins by processing it through DINOv2 to generate embeddings that capture its semantic information. These embeddings are then passed to the Retrieval Module, which retrieves the most similar images that are subsequently processed using a pre-trained image encoder, designed to handle multiscale features for robust image representation. This hierarchical structure enables the extraction of detailed and context-rich embeddings at multiple scales, which are essential for accurate segmentation.

Once the embeddings of the retrieved images are obtained, they are combined with their associated segmentation masks and input into the memory encoder. The memory encoder downsamples the segmentation masks and integrates them with the image embeddings generated by the image encoder, producing compact and efficient memory representations. These encoded memories are then stored in a memory bank, where they serve as the basis for the memory attention operation, facilitating precise and context-aware segmentation.

During the segmentation process, the input image is processed through the image encoder to generate its em-

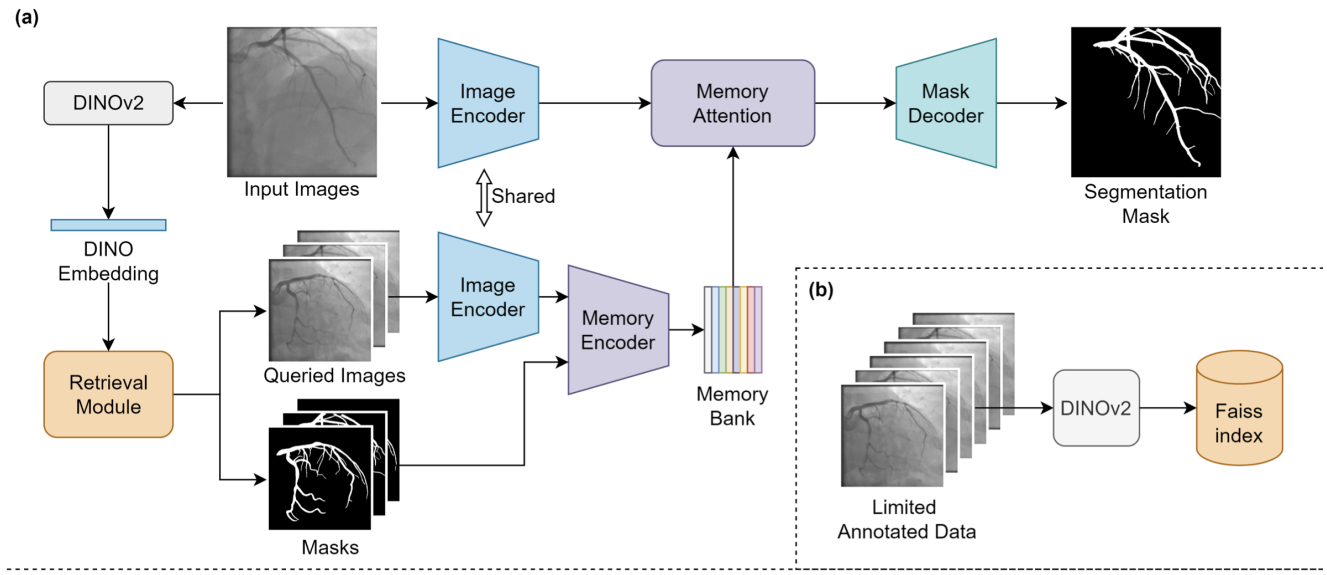

Figure 4. Overview of Retrieval-Augmented Few-Shot Medical Image Segmentation Workflow: (a) The main segmentation pipeline starts with input images processed through DINOv2 for dino embedding, followed by querying similar images and corresponding masks which are encoded and stored in a memory bank. The memory attention mechanism integrates the information from memory bank to assist the mask decoder in generating the final segmentation mask. (b) The process of indexing limited annotated data using DINOv2 and Faiss, enabling efficient retrieval of relevant images to enhance segmentation accuracy.

beddings. The model then employs the memory attention mechanism from SAM 2, which dynamically integrates information from the memory bank. This memory attention module consists of stacked transformer blocks that first apply self-attention to the current image features and then perform cross-attention with the stored memory representations. The enriched features produced by this process are subsequently passed to the mask decoder.

It is important to emphasize that our framework directly adopts the memory encoder, memory attention mechanism from SAM 2 without requiring any additional retraining or pretraining. During segmentation, the framework operates independently of external prompts for the mask decoder. Instead, the segmentation process is guided by the rich contextual information stored in the memory bank, which comprises encoded features from similar images.

This memory-driven approach enables the framework to accurately segment target structures by leveraging prior knowledge stored in the memory representations. As a result, the framework is highly efficient and effective, even when working with limited annotated data, making it particularly well-suited for medical imaging tasks.

4. Experiment

4.1. Dataset and Evaluation

For the ablation study on LoRA rank, tuning mode, and the impact of RAG inclusion or exclusion, we used the MICCAI FLARE 2022 challenge dataset. This dataset contains abdominal organ CT scans of 13 organs, with 2000 unlabeled cases and 50 labeled cases of pancreatic disease. The 2000 unlabeled cases are not for LoRA fine - tuning; we used the 50 - case labeled subset, splitting it into 30 scans for training and 20 for testing. However, comparing our methods with other segmentation methods on the MICCAI FLARE dataset is difficult because no other approaches were trained on this validation set. So, for a fairer and more reliable comparison, we used the BTCV dataset. The BTCV dataset has 30 abdominal CT scans with 13 - organ annotations. It's from the BTCV MICCAI Challenge, with 30 3D volumes divided into 24 for training and 6 for testing. In both the BTCV and FLARE datasets, CT scans are converted to 2D slices for SAM input. The BTCV dataset has 3,779 axial contrast - enhanced abdominal CT images in total, with 3,017 axial slices in the training set. Each CT volume has 85 - 198 slices, a resolution of 512×512 pixels, in - plane spatial resolution of ([0.54×0.54]˜[0.98×0.98]), and slice thickness of [2.5˜5.0]mm. The FLARE dataset has 2,857 training slices and 1,904 testing slices, each with a 512 × 512 - pixel resolution.

The effectiveness of the model is evaluated using the standard DSC (Dice Similarity Coefficient). The Mean Dice Coefficient is the average Dice score computed across multiple regions or classes within an image. It ranges from 0 (indicating no overlap) to 1 (indicating perfect overlap). This metric provides a comprehensive assessment of

the segmentation algorithm's performance across different structures in the image, making it particularly useful for multi-class or multi-region segmentation tasks.

The formula for the **Mean Dice Coefficient** is given as:

$$\text{Dice}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{3}$$

where X represents the predicted segmentation, and Y represents the ground truth segmentation.

The **Mean Dice Coefficient** is then calculated as the average Dice coefficient over all N regions or classes:

$$\text{Mean Dice} = \frac{1}{N} \sum_{i=1}^{N} \text{Dice}(X_i, Y_i) \tag{4}$$

where X_i and Y_i represent the predicted and ground truth segmentation for the i-th class or region, respectively.

4.2. Implementation details

For lora fintune version SAM: SAMed We set the LoRA rank to 4 and use the LoRA fine-tuning mode for the mask decoder, combined with RAG, for the final comparison against the SAMed "vit_b" version of SAM. Input images are resized to 512×512 before being fed into SAMed to maintain the predicted segmentation logits' resolution. There are 1 + n predicted segmentation logits, where n is the number of organ classes for the specific task, and the additional class represents the background. The loss function combines cross-entropy loss (weighted at 0.2) and Dice loss (weighted at 0.8). For the warmup process, we set the initial learning rate (Ilr) to 0.005, the warmup period (WP) to 250 iterations, and the maximum number of iterations to 18,600 (equivalent to 200 epochs). The AdamW optimizer is configured with $\beta 1 = 0.9$, $\beta 2 = 0.999$, and a weight decay of 0.1. We apply early stopping at 14,880 iterations (160 epochs) to ensure optimal training performance.

For the Adapter fintune version SAM: Medical-SAM-Adapter. we implemented the Med-SA pipeline primarily based on the official ViT-H SAM GitHub repository. For 2D medical image training, we followed the default training settings of SAM.

All the experiments are implemented with the PyTorch platform and trained/tested on 2 RTX 3090 GPUs.

4.3. Comparison with SOTAs

To evaluate the overall performance of our proposed fine-tuning models, we compare them with state-of-the-art (SOTA) segmentation methods on the multi-organ segmentation dataset BTCV. The quantitative results are presented in Table 1. Specifically, we assess the Medical-SAM-Adapter (Med-SA) and SAMed models alongside well-recognized medical image segmentation methods, including nnUNet [10], TransUNet [3], UNetr [8], and SegDiff [1]. Additionally, we include comparisons with the baseline SAM and the fully fine-tuned SAM (MedSAM) [17]. Segmentation performance is measured using the Dice score.

In Table 1, we observe that on the BTCV dataset, the Med-SA and SAMed achieve state-of-the-art (SOTA) performance across all 12 organs, surpassing other methods in overall performance.

The results demonstrate the effectiveness of LoRA and adapter fine-tuning. Med-SA and SAMed successfully bridge the gap between natural image segmentation and medical image segmentation under a unified framework, requiring customization of only a small fraction of the parameters in SAM. Compared to other methods, such as MedSAM model across all prompt variations, our approach outperforms MedSAM by updating only 2% of its total trainable parameters—13M for Med-SA and 6M for SAMed. This highlights the efficiency and effectiveness of the proposed techniques.

4.4. Ablation Study

In this section, we talk about how the performance of SAMed was enhanced via three ablation studies: LoRA rank, fine-tuning mode, and RAG inclusion or exclusion. For the ablation study, the SAMed model was trained on the MICCAI FLARE dataset, using the same dataset and implementation details mentioned previously. LoRA was applied to the q and v projection layers within the q, k, v, and o layers. The training strategy stayed the same, with fine-tuning on the image encoder, mask decoder, and prompt encoder, and the prompt input kept as a fixed embedding all through.
fintuning mode: The customization of the image encoder lets SAMed extract meaningful features from medical images, which greatly benefits the subsequent processing in the mask decoder. At first, we opted to fine-tune the mask decoder without using LoRA, enabling us to fine-tune the entire mask decoder directly. But later, we wondered if LoRA fine-tuning could enhance the mask decoder's performance further. This is because the mask decoder in SAM has a lightweight transformer layer for decoding the extracted image tokens. The updated parameter shape after applying LoRA is significantly smaller than the original SAMed, so we call this version SAMed_s. The performance comparison between SAMed and SAMed_s is shown in Table 2.

From the results, we notice that performance improves when the mask decoder is fine-tuned. We think this is due to the fact that the mask decoder already has the ability to handle subsequent processing, and LoRA fine-tuning keeps its original parameters, which is beneficial for medical imaging tasks. This approach makes sure the decoder maintains its original functionality while adapting well to the new task.

The illustration in Figure 5 further demonstrates that LoRA fine-tuning of the mask decoder improves organ segmentation. We present the test set results by visualizing the

Model	Spleen	R.Kid	L.Kid	Gall.	Eso.	Liver	Stom.	Aorta	IVC	Veins	Panc.	AG	Avg
TransUNet	0.952	0.927	0.929	0.662	0.757	0.969	0.889	0.920	0.833	0.791	0.775	0.637	0.838
SegDiff	0.954	0.932	0.926	0.738	0.763	0.953	**0.927**	0.846	0.833	0.796	0.782	0.723	0.847
UNetr	0.968	0.924	0.941	0.750	0.766	0.971	0.913	0.890	0.847	0.788	0.767	0.741	0.856
nnUNet	0.942	0.894	0.910	0.704	0.723	0.948	0.824	0.877	0.782	0.720	0.680	0.616	0.802
SAM 1 point	0.518	0.686	0.791	0.543	0.584	0.461	0.562	0.612	0.402	0.553	0.511	0.354	0.548
SAM 3 points	0.622	0.710	0.812	0.614	0.605	0.513	0.673	0.645	0.483	0.628	0.564	0.395	0.631
SAM BBox 0.5	0.346	0.585	0.592	0.375	0.426	0.377	0.451	0.536	0.392	0.576	0.426	0.202	0.420
MedSAM 1 point	0.751	0.814	0.885	0.766	0.721	0.901	0.855	0.872	0.746	0.771	0.760	0.705	0.803
MedSAM 3 points	0.758	0.831	0.889	0.782	0.733	0.917	0.858	0.876	0.755	0.776	0.763	0.716	0.820
MedSAM BBox 0.5	0.621	0.736	0.801	0.721	0.715	0.811	0.714	0.770	0.622	0.618	0.630	0.545	0.692
Med-SA (ours)	0.936	0.947	0.964	**0.842**	**0.788**	**0.983**	0.926	0.929	**0.879**	**0.852**	**0.790**	**0.823**	**0.889**
SAMed (ours)	**0.974**	**0.968**	**0.986**	0.830	0.770	0.931	0.924	**0.935**	0.869	0.774	0.788	0.689	0.870

Table 1. The comparison of Med-SA and SAMed with SOTA segmentation methods over BTCV dataset evaluated by Dice Score. Best results are denoted as **bold**.

Methods	DSC	Model size	Spleen	Kidney(R)	Kidney (L)	Gallbladder	Liver	Stomach	Aorta	Pancreas
SAMed	0.859	18.81M	0.939	0.850	0.806	**0.855**	0.953	0.858	**0.909**	0.698
SAMed_s	**0.868**	**6.33M**	**0.951**	**0.879**	**0.848**	0.813	**0.964**	**0.880**	0.895	**0.714**

Table 2. The ablation study of SAMed on the fine-tuning mode for the mask decoder on the FLARE dataset is evaluated using the Dice Score, with the better results marked in **bold**.

segmentation on the center slice of each scan. The comparison includes the outcomes from SAMed, SAMed_s, and the ground truth. It is evident that SAMed_s can segment certain organs more accurately, maintaining their correct sizes and appearing closer to the ground truth. In contrast, SAMed struggles to achieve this, as fine-tuning all parameters negatively impacts its segmentation ability.

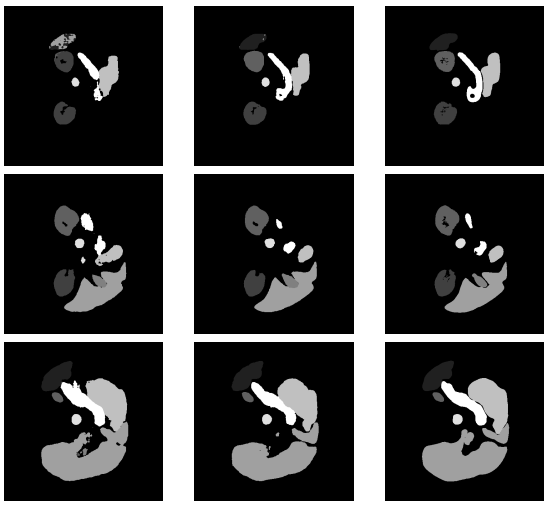

Figure 5. There are three examples from the FLARE 2022 test dataset, where the segmentations produced by SAMed and SAMed_s are compared against the ground truth.

Lora Rank LoRA rank is a crucial parameter in the context of Low-Rank Adaptation (LoRA). It refers to the dimensionality of the low-rank decomposition used in LoRA layers and determines the size of these low-rank matrices, which directly impacts the model's adaptability and performance. Table 3 presents the performance of SAMed as the rank size in the LoRA layers is adjusted.

We observe that a LoRA rank size of 4 gives the best performance, surpassing other rank sizes in segmenting six organs: Spleen, Left Kidney, Gallbladder, Liver, Stomach, and Pancreas. For other organs, the performance is also quite satisfactory. We find that the performance of SAMed gradually improves within a certain range of rank sizes but drops considerably when the rank becomes too large. We believe this is because the LoRA layers in SAMed need a sufficient number of parameters to effectively tailor the model for medical image datasets. However, an excessive number of trainable parameters may impede SAMed's ability to retain the original segmentation capabilities of SAM, thus increasing the complexity and difficulty of training.

The visual results further confirm our findings. Figure 6 illustrates the segmentation outcomes for LoRA rank sizes of 1, 4, and 16 on the middle slice of 3D scans. From these results, we observe that with a rank size of 1 or 16, some organs cannot be segmented correctly, leading to fragmented and noisy outputs that are unsuitable for practical applications. In contrast, a rank size of 4 produces accurate organ segmentations with complete size and shape. These find-

Rank size	DSC	Spleen	Kidney(R)	Kidney (L)	Gallbladder	Liver	Stomach	Aorta	Pancreas
1	0.859	0.939	0.850	0.806	0.855	0.953	0.858	**0.909**	0.698
4	**0.875**	**0.948**	0.875	**0.866**	**0.860**	**0.956**	**0.877**	0.903	**0.718**
16	0.809	0.925	**0.882**	0.863	0.719	0.940	0.757	0.812	0.571

Table 3. The ablation study of SAMed on lora rank size over FLARE dataset evaluated by Dice Score. Better results are denoted as **bold**.

W.o RAG	DSC	Spleen	Kidney(R)	Kidney (L)	Gallbladder	Liver	Stomach	Aorta	Pancreas
YES	**0.872**	**0.945**	**0.861**	**0.837**	**0.882**	0.940	**0.873**	0.895	**0.742**
NO	0.859	0.939	0.850	0.806	0.855	**0.953**	0.858	**0.909**	0.698

Table 4. The ablation study of SAMed on RAG over FLARE dataset evaluated by Dice Score. Better results are denoted as **bold**.

ings highlight the importance of setting the LoRA rank to 4 in the final experiments.

Figure 6. There are two examples from the FLARE 2022 test dataset, segmented using the model with LoRA rank sizes set to 1, 4, and 16, alongside the ground truth for comparison.

w.o RAG In this part, we aim to fully utilize the few-shot capabilities of LoRA. Inspired by LLMs, we integrate the RAG module into our framework. With RAG, the segmentation process is guided by the rich contextual information stored in the memory bank, which contains encoded features from similar images. As shown in Table 4, we find that the framework becomes highly efficient and effective, even with limited annotated data. This innovation slightly improves the model's ability to segment medical images, while all other settings remain unchanged.

From the results shown in Table 4, we observe that the Dice scores for nearly all organs have improved significantly with the integration of the RAG module. Notably, the overall Dice score shows an improvement of 1.3%, highlighting the effectiveness of RAG in enhancing the segmentation performance.

Furthermore, the visual results in Figure 7 provide compelling evidence of the positive impact of RAG on SAM. The figure illustrates that the integration of RAG enables SAM to produce more accurate and complete segmentations, particularly for challenging organs. This is especially advantageous when working with limited data or in cases where certain organs, such as the pancreas, are absent in some slices. By leveraging the contextual information stored in the memory bank, RAG allows the framework to generate precise and complete segmentations of the pancreas. In contrast, without RAG, the segmentation of the pancreas often appears fragmented and inconsistent. These findings underscore the critical role of RAG in enhancing segmentation quality in difficult scenarios, further reinforcing its value in medical image segmentation tasks.

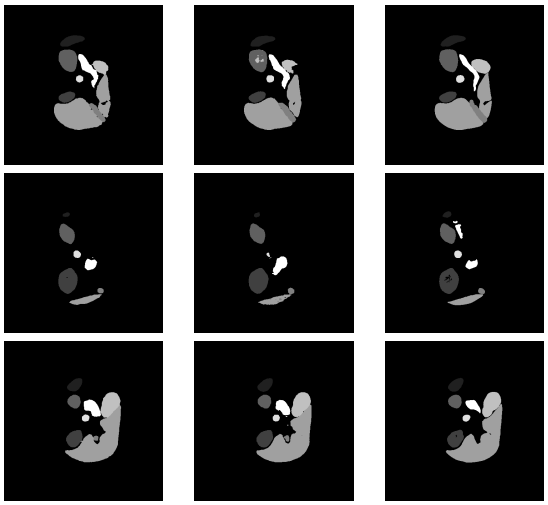

Figure 7. There are two examples from the FLARE 2022 test dataset, segmented using the model with RAG and without RAG, alongside the ground truth for comparison.

In conclusion, we enhanced the performance of SAMed by modifying the fine-tuning mode of the mask decoder, optimizing the core LoRA rank, and introducing the RAG module. These improvements significantly boosted the model's performance, bringing it close to the level of other state-of-the-art (SOTA) models.

References

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 6

[2] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis, 2022. 3

[3] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 6

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021. 2

[5] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 4

[6] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2024. 3

[7] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, 2021. 2

[8] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022. 6

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 4

[10] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 2021. 6

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

[12] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 1

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. 3

[14] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge, 2023. 3

[15] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition, 2022. 1, 3

[16] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, Fan Zhang, Wentao Liu, YuanKe Pan, Shoujin Huang, Jiacheng Wang, Mingze Sun, Weixin Xu, Dengqiang Jia, Jae Won Choi, Natália Alves, Bram de Wilde, Gregor Koehler, Yajun Wu, Manuel Wiesenfarth, Qiongjie Zhu, Guoqiang Dong, Jian He, the FLARE Challenge Consortium, and Bo Wang. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge, 2023. 1

[17] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 2024. 6

[18] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 4

[19] Junde Wu, Wei Ji, Huazhu Fu, Min Xu, Yueming Jin, and Yanwu Xu. Medsegdiff-v2: Diffusion based medical image segmentation with transformer, 2023. 2

[20] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation, 2023. 1, 4

[21] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation, 2023. 1, 3