RevSAM2: Prompt SAM2 for Medical Image Segmentation via Reverse-Propagation without Fine-tuning

Yunhao Bai[1]   Boxiang Yun[1]   Zeli Chen[2]   Qinji Yu[3]   Yingda Xia[2]   Yan Wang[1*]

[1]Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University

[2]DAMO Academy, Alibaba Group

[3]Shanghai Jiao Tong University, Shanghai, China.

Abstract

The Segment Anything Model 2 (SAM2) has recently demonstrated exceptional performance in zero-shot prompt segmentation for natural images and videos. However, when the propagation mechanism of SAM2 is applied to medical images, it often results in spatial inconsistencies, leading to significantly different segmentation outcomes for very similar images. In this paper, we introduce RevSAM2, a simple yet effective self-correction framework that enables SAM2 to achieve superior performance in unseen 3D medical image segmentation tasks without the need for fine-tuning. Specifically, to segment a 3D query volume using a limited number of support image-label pairs that define a new segmentation task, we propose reverse propagation strategy as a query information selection mechanism. Instead of simply maintaining a first-in-first-out (FIFO) queue of memories to predict query slices sequentially, reverse propagation selects high-quality query information by leveraging support images to evaluate the quality of each predicted query slice mask. The selected high-quality masks are then used as prompts to propagate across the entire query volume, thereby enhancing generalization to unseen tasks. Notably, we are the first to explore the potential of SAM2 in label-efficient medical image segmentation without fine-tuning. Compared to fine-tuning on large labeled datasets, the label-efficient scenario provides a cost-effective alternative for medical segmentation tasks, particularly for rare diseases or when dealing with unseen classes. Experiments on four public datasets demonstrate the superiority of RevSAM2 in scenarios with limited labels, surpassing state-of-the-arts by 12.18% in Dice. The code will be released.

1. Introduction

The Segment Anything Model 2 (SAM2) [28] has shown remarkable zero-shot prompt segmentation capabilities in natural images and videos. With points, boxes or mask
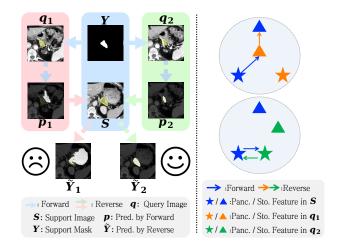

Figure 1. Example of reverse propagation (left) and in feature space (right). S is a CT slice and Y is its segmentation mask, while q_1 and q_2 are two adjacent CT slices from a different CT scan than S. The pancreatic tail region in all three images is outlined in yellow. The prediction masks p_1 and p_2 correspond to q_1 and q_2, respectively, generated by the memory bank that stores features of S and Y. Conversely, $\tilde{Y}_1(\tilde{Y}_2)$ is the prediction masks for S generated by the memory bank filled with the features of p_1 and q_1 (p_2 and q_2). In our framework, p_1 will be discarded, and q_2 along with p_2 are used to support the re-segmentation of q_1.

as prompts, SAM2 can accurately segment the object foreground within an image and track the object across frames in a video. However, similar to its predecessor, SAM [20], SAM2 struggles with medical images, particularly in Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans. This limitation stems from the lack of medical images in its training data, resulting in the model's inability to precisely delineate the foreground of organs and other structures in medical images via their semantic features.

Recent studies [3, 26, 41] have sought to adapt SAM2 for medical imaging applications. Like fine-tuning on SAM, these approaches typically involve fine-tuning some components of SAM2, such as the mask decoder, using a certain amount of labeled data. However, these methods face two

primary limitations: first, they require a substantial amount of labeled data and considerable time for training, which cannot handle unseen tasks or settings under domain shift; second, even after training, they still depend on interactive prompts to perform segmentation on the target images. Moreover, only taking cues from SAM, these methods inevitably overlook the fact that compared with SAM, SAM2 has made meaningful enhancements in its network architecture, resulting in a more capable and versatile successor. SAM2 uses a memory attention module to access information stored in the memory bank about the target object, enabling it to segment the corresponding target in subsequent images. This inspires us to explore it from a different perspective, *i.e.*, is it possible to leverage SAM2's ability to automatically segment targets in medical images such as CT and MRI (referred to as query images) based on only a few similar images, specifically a few 2D slices (called support images), and their labels, without *any* fine-tuning? This setting shows promising research potential in many clinical scenarios, such as in cases where clinical researchers often lack the resources and expertise to train neural networks [1], only a few or scarce labels can be provided for certain rare diseases or working with unseen classes.

Compared to training a model from scratch [13, 16] using a few labeled slices, fine-tuning a foundation model [6, 17, 19, 38] with such data seems more feasible. However, the extremely limited amount of data still poses a significant challenge. Some few-shot methods [5, 35, 36, 42] attempt to perform inference in scenarios with limited unseen class data, but these methods still require a large number of data from other classes to train a support-query segmentation model.

In this paper, we fully leverage the memory bank and propagation ability of SAM2. Since SAM2 is trained on natural images and videos, giving it the ability to locate objects in target images even when facing certain positional and appearance changes. A straightforward approach to segment 3D medical images based on a few labeled support images using SAM2 is to store the support images and their masks in the memory bank, and then maintain a first-in-first-out (FIFO) queue of memories to predict each query slice. However, this approach may lead to inferior performance if previous query slices are not well segmented. This is because, when applied to unseen medical images, SAM2 often overemphasizes the positional and appearance information of target tissues stored in the memory bank. As a result, incorrect segmentation occurs when target tissues experience positional or appearance shifts, especially if the stored information is not generalizable. Although tissues in CT and MRI images, such as organs, tend to share similar locations and appearances among patients, considerable variations exist due to patient heterogeneity, making segmentation challenging *without* fine-tuning. This observa-

tion led us to consider whether it is possible to automatically select well-segmented query slices (conditional slices) and propagate them to slices without prompts (non-conditional slices) *within the same query volume*.

Without any groundtruth masks, deciding whether a query slice is well segmented seems challenging, since the IoU score predictor in SAM2 trained specifically on natural images is not reliable for evaluating medical images. To solve this problem, we propose a **surprisingly simple yet effective** reverse-propagation strategy. Concretely, we first forward propagate the features of support images and masks to obtain the prediction for each slice in the query volume. Then, we reverse-propagate the query image along with its predicted mask back to derive masks for the support images. The quality of the query slice prediction is evaluated using the Dice scores between the predicted masks and the ground-truth masks of support images. Fig. 1 (left) illustrates the motivation of our simple process. As shown in Fig. 1, features of S and Y propagates an incorrect prediction p_1 for a query slice q_1 but a correct prediction p_2 for its adjacent query slice q_2. After reverse propagation, q_1 and p_1 lead to an even more incorrect prediction $\widetilde{Y}_1$ in S, while q_2 and p_2 yield an accurate prediction $\widetilde{Y}_2$. This intriguing phenomenon can be explained in Fig .1 (right) in the feature space: for the query slice q_1, the feature of the real target pancreas (★) may encounter some positional or appearance shift which deviates from the pancreas' feature of the support image stored in the memory (★) by a certain distance. This makes SAM2 to find a feature of another tissue stomach (▲) closer to the pancreas' feature (★). During reverse propagation, the predicted target in the q_1 (▲) has a higher chance to find another closer tissue *e.g.*, stomach (▲) in the support image, rather than pancreas (★). But for the query slice q_2, since the distance between the feature of the pancreas in the support image (★) and the one in q_2 (★) is small, there is a high probability that after reverse propagation, the pancreas (★) can still be identified.

Based on the proposed reverse-propagation, we design RevSAM2, a self-correction framework, which prompts SAM2 for medical image segmentation via reverse-propagation without fine-tuning. Given only a few 2D support images and their corresponding segmentation masks, our framework can segment the query volume without additional prompts or retraining, adapting quickly by simply changing the support images. We validate RevSAM2 on four publicly available multiorgan datasets, demonstrating its effectiveness by achieving state-of-the-art (SOTA) performance on all datasets. Furthermore, we also validated the robustness of RevSAM2 for domain adaptation in scenarios with limited labels.

2. Related Work

2.1. Segment Anything Model 2

Compared to Segment Anything Model [20] (SAM), which focuses solely on promptable image segmentation, Segment Anything Model 2 (SAM2) introduces an additional capability for promptable video segmentation. The components responsible for image segmentation in SAM2 remain the same as those in SAM: the image encoder, prompt encoder, and mask decoder. In SAM2, the image encoder and prompt encoder independently encode the input image and prompt information, which are then fused and passed into the mask decoder to generate the segmentation mask.

For video segmentation, SAM2 incorporates additional components: memory encoder, memory bank, and memory attention. The workflow is as follows: First, **add prompt**. Prompts are applied to select frames (conditional frames), which undergo independent image segmentation to produce masks. Second, **fill memory bank**. The memory encoder encodes the features of images and masks, storing them in the memory bank. Finally, **propagate**. For non-conditional frames, image features are extracted and processed with memory attention using the features stored in memory bank. After generating the mask, step two is repeated.

2.2. SAM-based Medical Image Methods

Segment Anything Model [20] (SAM) has been extensively trained on over a billion natural images, demonstrating strong zero-shot segmentation capabilities for natural images: given a prompt (e.g., a point or a bounding box) for an image, it can segment the object foreground indicated by the prompt. Some previous works [2, 7, 40] attempted to apply SAM to medical images, but experiments [4, 8, 14, 30] have shown that SAM's zero-shot segmentation performance significantly drops on unseen medical images. Therefore, recent studies have focused on how to fine-tune SAM using a certain amount of medical images to adapt it to medical images [6, 17, 19, 37, 38]. For instance, MedSAM [25] created a large-scale medical image dataset to retrain SAM with bounding box prompts; SAMed [17] introduced LoRA [15] layers into the image encoder while using the original mask decoder; H-SAM [6] used LoRA-inserted image encoders to enhance SAM's feature extraction capability for medical images and designed a hierarchical mask decoder to enable prompt-free segmentation.

Similar to SAM, SAM2 also performs suboptimally on medical images when using point and box prompts. After SAM2 was released, several studies built upon the successful adaptations of SAM for medical imaging, aiming to adapt SAM2 for the medical image domain. For example, MedicalSAM2 [41] fine-tune the all components of SAM2 exclude prompt encoder, MedSAM [26] only fine-tune the image encoder and mask decoder, while SAM2-

Adapter [3] introduces lightweight adapters into the image encoder, which are fine-tuned alongside the mask decoder during weight updates, using 4*A100 GPUs for training.

In the SAM2 architecture, the image encoder is significantly more complex than the mask decoder. The success of these approaches, which primarily focus on fine-tuning the mask decoder, suggests that SAM2's image encoder is already capable of effectively encoding the information present in medical images. In contrast to these methods, our approach aims to adapt SAM2 to medical imaging in a more challenging setting: insufficient label and without any weight fine-tuning.

2.3. Label Insufficient Medical Image Segmentation

Label Insufficient Medical Image Segmentation is an emerging field that addresses the challenge of limited annotated data—a common issue in the medical domain due to the high costs, complexity of annotation, and legal constraints on data sharing. Specifically, a major research focus within this field is Few-Shot Medical Image Segmentation (FSMIS). FSMIS techniques can be broadly categorized into two main types: prototypical network-based models [12, 27, 31, 34, 39] and two-branch interaction-based models [9, 10, 29, 33]. These methods train a model to segment the query image by drawing information from the support image and its mask, and then validate the model on unseen categories that are not part of the training process.

However, most of these methods treat consecutive slices as independent entities, segmenting each slice separately in a few-shot manner. We believe that in a sequence of consecutive slices, the segmentation results of earlier slices can be leveraged to assist in the segmentation of subsequent slices.

3. Method

Mathematically, we first define a 3D volume as X, and its ith slice is denoted as x_i. Given the 3D volume of a query medical image as $Q \in \mathbb{R}^{M \times W \times H}$, a set of few 2D support images as $S \in \mathbb{R}^{N \times W \times H}$ in the 3D form, and their corresponding segmentation labels as $Y \in \{0, 1\}^{N \times W \times H}$, which indicate the background and target regions in S. Our goal is to predict the per-voxel segmentation map $\widetilde{P} \in \mathbb{R}^{M \times W \times H}$ of Q using the few support image-label pairs. Our segmentation framework, RevSAM2, is built upon SAM2.

The overall pipeline of the proposed RevSAM2 is illustrated in Fig. 2 (a). For the ith slice q_i in volume Q, we first obtain its prediction p_i by using the memory bank which stores the features of S and Y through memory attention mechanism. We then assess the quality of p_i using reverse propagation, which treats q_i-p_i pair as a support image-label pair, and predict segmentation map $\widetilde{Y}_i$ for all support 2D images S. After obtaining the average predicted Dice scores for S, we select the top k images with the highest scores. These retained images, along with their corresponding predictions, serve as conditional images to propagate
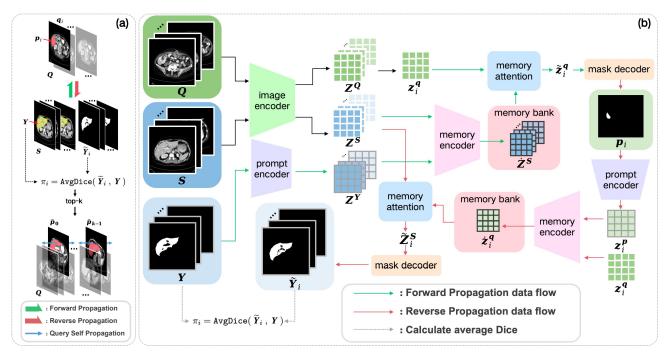
Figure 2. The overall framework of RevSAM2 (a) and illustration of forward propagation and reverse propagation (b). To evaluate the quality of the prediction p_i obtained by forward propagating S and Y onto q_i, we reverse propagate q_i and p_i back to S to obtain $\widetilde{Y}_i$, and calculate the average dice π_i between $\widetilde{Y}_i$ and Y and treat it as the as the metric to evaluate the accuracy of p_i.

information to all slices within Q through self-propagation, ultimately resulting in prediction $\widetilde{P}$.

3.1. Forward Propagation

As illustrated in Fig. 2 (b), we first use an image encoder $\mathcal{E}^{\mathrm{I}}$, parameterized by Θ^{I}, and a prompt encoder $\mathcal{E}^{\mathrm{P}}$, parameterized by Θ^{P} to encode S and Y, respectively, to obtain image features Z^S and prompt features Z^Y:

$$Z^S = \mathcal{E}^{\mathrm{I}}(S, \Theta^{\mathrm{I}}), \qquad (1)$$
$$Z^Y = \mathcal{E}^{\mathrm{P}}(Y, \Theta^{\mathrm{P}}). \qquad (2)$$

After feature extracting of support images and their segmentation labels, memory encoder sums Z^Y element-wise with Z^S, followed by a lightweight convolutional layer, whose parameters are Θ^{M}, to fuse the information and obtain $\dot{Z}^S$:

$$\dot{Z}^S = \mathcal{E}^{\mathrm{M}}(Z^S, Z^Y, \Theta^{\mathrm{M}}), \qquad (3)$$

where $\mathcal{E}^{\mathrm{M}}(\cdot, \cdot, \Theta^{\mathrm{M}})$ means the memory encoder.

The memory bank retains information about the segmentation of S for the target object segmentation in Q. Specifically, for the ith slice q_i in Q ($0 \leq i < M$), its image feature embedding z_i^q is extracted by the image encoder as described for S in Eq. 1: $z_i^q = \mathcal{E}^{\mathrm{I}}(q_i, \Theta^{\mathrm{I}})$. Then, conditioning z_i^q on $\dot{Z}^S$ through self-attention and cross-attention in the memory attention mechanism, we obtain the fused vision feature $\widetilde{z}_i^q$, which is formulated as:

$$\widetilde{z}_i^q = \mathcal{A}(z_i^q \mid \dot{Z}^S, \Omega), \qquad (4)$$

where $\mathcal{A}(\cdot \mid \cdot, \Omega)$ means the memory attention module whose parameters are Ω. Finally, $\widetilde{z}_i^q$ fed into the mask decoder $\mathcal{D}^{\mathrm{M}}$ to obtain the predicted segmentation map p_i:

$$p_i = \mathcal{D}^{\mathrm{M}}(\widetilde{z}_i^q, \Gamma^{\mathrm{M}}), \qquad (5)$$

where Γ^{M} indicate the parameter of the mask decoder.

Thus, we obtain the segmentation mask p_i of q_i based on the information from S and Y, without any direct prompt from q_i. However, due to the spatial positional perturbation of SAM2 mentioned in Sec. 1, the quality of p_i may not be optimal. Next, we propose a novel reverse-propagate strategy to evaluate the quality of p_i by checking whether it can reverse propagate back to S and generate Y.

3.2. Reverse Propagation

After obtaining p_i, we apply reverse propagation to calculate the predicted Dice scores for S using q_i-p_i pair as a support image-label pair. This simple yet effective evaluation strategy can determine whether p_i is accurate or affected by positional perturbation. Similar to how we obtain $\dot{Z}^S$ from S and Y, we first derive $\dot{z}_i^q$ from q_i and p_i as follows:

$$z_i^p = \mathcal{E}^{\mathrm{P}}(p_i, \Theta^{\mathrm{P}}), \qquad (6)$$
$$\dot{z}_i^q = \mathcal{E}^{\mathrm{M}}(z_i^q, z_i^p, \Theta^{\mathrm{M}}). \qquad (7)$$

The resulting $\dot{z}_i^q$ is then used to reversely segment S. Similarly, the image features Z^S of S are processed through
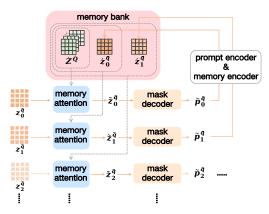
Figure 3. The illustration of the query self propagation. In query self propagation, the memory bank continuously stores the features of conditional slices selected by reverse propagation, while maintaining a FIFO queue to store the features of non-conditional slices during internal query inference.

the memory attention module, combined with $\dot{z}_i^q$, to generate the fused vision features $\widetilde{Z}_i^S$:

$$\widetilde{Z}_i^S = \mathcal{A}(Z^S \mid \dot{z}_i^q , \Omega). \tag{8}$$

Note that the ith q_i-p_i pair yields $\widetilde{Z}_i^S$. The prediction map $\widetilde{Y}_i$ is then obtained by feeding $\widetilde{Z}_i^S$ to the mask decoder:

$$\widetilde{Y}_i = \mathcal{D}^{\mathtt{M}}(\widetilde{Z}_i^S , \Gamma^{\mathtt{M}}). \tag{9}$$

The purpose of reverse propagation is to amplify the influence of wrong direction of propagation. If p_i is a correct segmentation, it is likely to support the accurate segmentation of S. Conversely, an incorrect segmentation of q_i may lead to significant errors in the segmentation of S, resulting in predictions that deviate substantially from Y. Thus, we calculate the average dice score between $\widetilde{Y}_i$ and Y, as a metric to evaluate the accuracy of p_i:

$$\pi_i = \mathtt{AvgDice}(\widetilde{Y}_i , Y). \tag{10}$$

After obtaining all scores of P, we retain the top k scoring masks among all p_i in P to get the final volume prediction map for Q by query self propagation (described in Sec. 3.3)

Intuitively, when there are more support images, the feature bias of each support image become smaller, making the response Dice obtained through reverse propagation more representative and accurate. This aspect will be further discussed in the ablation study.

3.3. Query Self Propagation

The overall framework of query self propagation is illustrated in Fig. 3. We refer to the query slices with the top k scoring prediction masks as conditional query slices, denoted as $\hat{Q}$, where $\hat{Q} \in \mathbb{R}^{k \times W \times H}$, and their corresponding prediction masks as $\hat{P}$. We set $k = 7$ as default .The

remaining query slices are referred to as non-conditional query slices, whose 3D volume is denoted as $\bar{Q}$.

We first extract the image features $Z^{\hat{Q}}$ and prompt features $Z^{\hat{P}}$, and then combine them to obtain the feature $\dot{Z}^{\hat{Q}}$ stored in the memory bank, similar to $\dot{Z}^S$ in propagation and $\dot{z}_i^q$ in reverse propagation:

$$Z^{\hat{Q}} = \mathcal{E}^{\mathtt{I}}(\hat{Q} , \Theta^{\mathtt{I}}), \tag{11}$$

$$Z^{\hat{P}} = \mathcal{E}^{\mathtt{P}}(\hat{P} , \Theta^{\mathtt{P}}), \tag{12}$$

$$\dot{Z}^{\hat{Q}} = \mathcal{E}^{\mathtt{M}}(Z^{\hat{Q}} , Z^{\hat{P}} , \Theta^{\mathtt{M}}). \tag{13}$$

Unlike the static memory bank used in the forward propagation and reverse propagation, the memory bank in query self propagation not only retains information about the conditional query slices but also maintains a first-in-first-out (FIFO) queue of memories of the τ most recent non-conditional query slices and their predicted masks. Here, we follow the default configuration $\tau = 7$ of SAM2 TINY Model. For each non-conditional query slice $\bar{q}_i$, its image feature $z_i^{\bar{q}}$ is extracted by image encoder: $z_i^{\bar{q}} = \mathcal{E}^{\mathtt{I}}(\bar{q}_i , \Theta^{\mathtt{I}})$. Then, we obtain the fused vision feature of the jth non-conditional query slice $\bar{q}_j$ via:

$$\widetilde{z}_j^{\bar{q}} = \mathcal{A}(z_j^{\bar{q}} \mid \dot{Z}^{\hat{Q}} , \dot{z}_{j-1}^{\bar{q}}, \ldots, \dot{z}_{\max(j-\tau,0)}^{\bar{q}}, \Omega). \tag{14}$$

We use recent non-conditional query slices to reduce the feature bias, as the positional and appearance shifts between conditional and non-conditional slices are smaller than those between support images and non-conditional slices. Finally, we obtain the prediction mask $\widetilde{p}_j^{\bar{q}}$ for $\bar{q}_j$:

$$\widetilde{p}_j^{\bar{q}} = \mathcal{D}^{\mathtt{M}}(\widetilde{z}_j^{\bar{q}} , \Gamma^{\mathtt{M}}). \tag{15}$$

All non-conditional prediction masks $\widetilde{P}^{\bar{Q}}$ are combined with the conditional masks $\hat{P}$ to form the final volume prediction mask $\widetilde{P}$ for the query Q.

4. Experiments

4.1. Datasets and evaluation

In order to validate our method, we conduct experiments on each single-organ segmentation task across four multi-organ medical image segmentation datasets, including BTCV [21], AbdomenCT-1K [24], Synapse-CT [22] and CHAOS-MRI [18].

BTCV The Multi-Atlas Labeling Beyond the Cranial Vault (BTCV) challenge dataset contains 50 abdominal CT scans. Initially, the dataset includes annotations for 13 organs, with additional annotations for the duodenum added in [11]. For our study, we use the same data split and 14 multi-organ labels as in [23].

AbdomenCT-1K This dataset consists of 1112 CT scans from five different sources, with annotations for the liver,

Method	spleen	kidnetR	kidneyL	gall	eso	liver	stomach	arota	IVC	veins	pancreas	AG R	AG L	duode	mDSC	mNSD
nnU-Net [16] (ALL)	96.14	88.19	81.15	84.86	78.14	90.41	85.05	91.32	87.92	70.17	77.87	73.32	75.15	77.95	82.70	84.40
Swin UNETR [13] (ALL)	95.69	87.80	95.31	80.60	80.37	89.39	76.66	90.21	86.85	66.24	74.71	72.68	75.80	71.30	81.69	83.27
1pos 1neg points	56.55	60.39	63.07	16.44	6.40	60.90	45.14	50.46	17.58	10.71	10.57	2.35	1.21	12.26	29.57	13.53
3 pos 3 neg points	62.73	73.87	73.98	15.77	12.52	68.82	47.38	48.34	17.06	7.72	11.39	1.75	1.82	11.47	32.47	15.40
bbox	86.41	81.91	80.56	64.04	34.18	76.86	47.89	39.47	38.24	33.62	31.31	25.10	22.77	36.54	49.92	40.56
mask	91.67	87.72	81.37	66.83	36.36	79.71	50.39	87.45	69.27	47.72	28.84	26.47	38.81	45.54	59.87	56.16
nnU-Net [16]	18.00	25.40	30.60	7.66	10.27	32.49	12.33	35.85	14.06	11.31	18.46	4.75	11.91	8.47	17.27	16.18
Swin UNETR [13]	20.90	19.92	8.40	6.32	14.79	33.56	11.81	28.38	16.15	8.23	15.92	7.31	21.95	11.45	16.08	13.58
UniverSeg [1]	58.95	65.25	64.40	24.43	22.82	77.10	41.61	47.03	35.86	27.62	21.75	15.56	12.27	17.78	38.03	45.34
SAMed [17]	71.59	47.11	52.64	31.83	32.36	84.03	40.30	56.99	36.39	37.24	24.96	4.99	8.04	15.42	39.49	33.85
H-SAM [6]	87.93	74.74	81.95	55.75	45.42	89.70	55.05	80.09	52.81	42.68	28.80	29.97	31.72	26.51	55.94	50.60
HQ-SAM [19]	74.54	69.54	73.97	60.35	74.05	67.41	63.38	62.97	75.46	14.14	46.97	36.01	51.80	40.97	57.97	48.41
CAT-SAM [38]	72.53	52.26	60.73	58.97	51.13	87.33	75.08	63.62	65.06	61.52	59.93	30.86	41.57	57.08	59.83	55.46
MedicalSAM2 [41]	78.66	73.90	69.85	56.13	45.08	85.49	61.63	30.18	50.49	27.32	45.75	33.03	23.42	46.40	51.95	38.61
RevSAM2(ours)	93.99	82.50	85.78	81.19	55.42	92.85	70.70	86.07	78.21	54.91	57.96	50.29	45.20	42.97	69.86	67.66

Table 1. Comparison of mDSC (%) and mNSD (%) on the BTCV dataset. 'ALL' refers to using the full training data to demonstrate the upper bound. For each category, only 10 slices are used as the training set or support images, with all methods using the same slices for comparison. The best performances are highlighted in **bold**, while the second-best performances are indicated with underlines.

kidney, spleen, and pancreas. [32] extended the label set to 13 organs. Like BTCV, we adopt the same data split and 13 multi-organ labels as used in [23].

Synapse-CT Synapse-CT is an abdominal CT dataset acquired from the MICCAI 2015 Multi-Atlas Abdomen Labeling Challenge, which comprises 30 3D abdominal CT scans. Following GMRD [5], we evaluate RevSAM2 on left kidney, right kidney, liver, and spleen.

CHAOS-MRI CHAOS-MRI is an abdominal MRI dataset obtained from the ISBI 2019 Combined Healthy Abdominal Organ Segmentation Challenge. It contains 20 3D T2-SPIR MRI scans. Following GMRD [5], we also evaluate RevSAM2 on the left kidney, right kidney, liver and spleen.

We conduct individual segmentation experiments for each class in the datasets. Unless otherwise specified, we randomly select 10 separated slices from three volumes in the training set for each class to simulate a real-world label insufficient scenario. To mitigate the effect of random selection, we use three groups of random slices, and report the average performance. The details of each group are provided in the supplementary material. The final evaluation metrics for our method are mean Dice Similarity Coefficient (mDSC) and mean Normalized Surface Dice (mNSD).

4.2. Comparison Methods

Our approach employs the SAM2 TINY model, which is the most lightweight variant. Initially, we test its prompt segmentation capabilities on medical images using points (1 positive and 1 negative, 3 positive and 3 negative), bounding box (bbox), and mask. For point prompts, we randomly select pixels from the ground truth as positive points and chose negative points from the false positives generated by the segmentation of positive points, repeating this process for each slice. Regarding bounding boxes and masks, we adhere to SAM2's evaluation method: generating bounding boxes based on the ground truth of the first slice or using the ground truth directly as mask prompts, allowing automatic propagation to obtain volume predictions.

We then conduct comparative experiments with four different approaches. The **first** type of approach involves training from scratch, such as nnU-Net [16] and Swin UN-ETR [13]. As expected, with only 10 labeled slices, these methods struggled to train models with good generalization. We also train them using all available training data to demonstrate their upper bounds. The **second** approach is a universal few-shot segmentation model, UniverSeg[1], which has been trained on a large number of medical image datasets to create a support-query style segmentation model. However, due to its limitation to 128x128 resolution input images, we have to compress the images for comparison. The **third** type of approach indicate SAM-based fine-tuning, including SAMed [17], H-SAM [6], HQ-SAM [19], CAT-SAM [38], and MedicalSAM2 [41]. These methods leverage foundation models (SAM [20] or SAM2 [28]) that have been extensively trained on natural images, making it more likely to fine-tune a segmentation model with some level of generalization in extremely data-scarce scenarios (10 slices). Notably, only SAMed and H-SAM can perform fully automatic segmentation, while other SAM-based methods still require prompts for query slices. **Finally**, we compare our results with traditional few-shot methods, including AAS_DCL [36], SR&CL [35], RPT [42] and GMRD [5], which train on some categories within the training set and test on unseen categories in the test set. It is worth noting that for these comparisons, we strictly follow their one-shot strategy, employing their exact support-query selection method. More details are illustrated in supplementary material.

4.3. Main Results

BTCV and AbdomenCT Table 1 and Table 2 show the results on BTCV dataset and AbdomenCT dataset, respectively. RevSAM2 shows outstanding segmentation ability with insufficiency labels (10 slices for each organ) on both datasets. As shown in the tables, both the re-trained models (nnUNet, Swin UNETR) and the SAM-based fine-tuning

Method	liver	kidneyR	spleen	pancreas	arota	IVC	AGL	AGR	gall	eso	stomach	duode	kidneyL	mDSC	mNSD
nnU-Net [16] (ALL)	99.03	98.29	99.01	90.46	97.65	95.16	90.64	89.79	93.05	90.11	97.38	91.85	98.17	94.70	95.70
Swin UNETR [13] (ALL)	98.95	98.18	98.92	88.26	97.45	95.38	90.45	89.35	92.22	88.47	96.49	88.78	98.30	94.00	94.50
1pos 1neg points	77.25	85.89	79.18	17.19	51.02	12.24	1.00	4.18	15.25	6.06	35.47	20.06	78.79	37.20	18.93
3 pos 3 neg points	79.54	95.37	90.62	12.45	57.83	20.14	1.72	2.36	12.34	11.04	43.58	15.13	89.30	40.88	24.34
bbox	89.60	82.48	92.01	53.56	33.72	54.16	25.75	26.83	45.75	22.25	64.07	41.52	84.94	55.13	37.48
mask	90.00	95.30	95.62	36.05	91.42	79.19	27.41	54.39	62.06	70.55	62.81	34.67	94.01	68.73	61.42
nnU-Net [16]	71.52	76.22	46.93	24.36	61.35	57.44	21.21	31.95	12.90	25.73	11.82	11.53	68.00	40.64	29.78
Swin UNETR [13]	53.84	43.18	53.44	21.88	41.58	31.68	28.29	15.62	14.02	23.03	12.22	5.94	48.48	30.25	23.35
UniverSeg [1]	89.51	77.86	77.87	29.29	58.60	50.16	23.49	29.01	38.98	44.78	56.81	33.57	77.03	52.84	58.48
SAMed [17]	89.51	74.36	69.81	38.03	68.86	35.51	35.71	11.37	60.50	56.67	48.66	25.49	79.30	53.37	39.20
H-SAM [6]	90.32	93.76	91.15	51.25	79.41	69.98	54.76	42.43	71.07	59.16	74.30	44.06	92.00	70.28	58.68
HQ-SAM [19]	71.17	68.96	71.04	39.54	69.10	84.76	42.65	8.86	67.54	67.91	70.31	46.71	80.23	60.68	42.55
CAT-SAM [38]	84.91	82.17	80.37	66.76	67.58	65.14	41.96	47.05	67.81	59.62	81.18	61.22	80.93	68.21	56.28
MedicalSAM2 [41]	95.18	76.09	90.57	59.60	29.20	54.96	43.28	40.49	79.35	62.16	74.68	59.14	79.38	64.93	44.06
RevSAM2(ours)	95.49	96.23	97.07	68.92	93.16	81.78	67.65	71.53	83.80	75.32	90.74	54.85	95.48	82.46	77.33

Table 2. Comparison of mDSC (%) and mNSD (%) on the AbdomenCT-1K dataset with other methods, only 10 slices are used as the training set or support images.

Query Dataset	BTCV		AbdomenCT	
Support Dataset	BTCV	Abd.	Abd.	BTCV
UniverSeg[1]	38.83	44.53	52.84	52.49
SAMed [17]	39.67	40.74	53.37	44.76
H-SAM [6]	56.96	55.04	70.28	63.04
HQ-SAM [19]	61.34	59.52	60.68	59.39
CAT-SAM [38]	59.70	59.50	68.21	68.10
MedicalSAM2 [41]	53.85	54.10	64.93	65.26
Ours	71.01	69.03	82.46	77.47

Table 3. Domain adaptation mDSC (%) comparison results on the BTCV and AbdomenCT datasets. Query Dataset: the dataset of test images; Support Dataset: the dataset of support images.

Method	Dataset	spleen	liver	LK	RK	mean
AAS-DCL [36]	Synapse-CT	72.30	78.04	74.58	73.19	74.52
SR&CL [35]		73.41	76.06	73.45	71.22	73.53
RPT [42]		79.13	82.57	77.05	72.58	77.83
GMRD [5]		78.31	79.60	81.70	74.46	78.52
Ours		94.02	89.41	84.20	82.05	87.42
AAS_DCL [36]	CHAOS-MRI	76.24	72.33	80.37	86.11	78.76
SR&CL [35]		76.01	80.23	79.34	87.42	80.77
RPT [42]		76.37	82.86	80.72	89.82	82.44
GMRD [5]		76.09	81.42	83.96	90.12	82.90
Ours		85.82	88.44	81.63	85.22	85.28

Table 4. Comparison of mDSC (%) with other few-shot methods on Synapse-CT (top) and CHAOS-MRI (bottom). We test the one-shot setting of these methods, strictly following their support-query pair selection strategies.

methods fail to achieve satisfactory segmentation performance, even for those using bounding box prompts (HQ-SAM, CAT-SAM, and MedicalSAM2) per slice. This is due to the extreme scarcity of labeled data, preventing these methods from converging to a model with good generalization. For UniverSeg, as it can only handle low-resolution images of 128x128, it performs well on larger organs (e.g., liver: 77.10% on BTCV, 89.51% on AbdomenCT) but poorly on smaller organs (e.g., right adrenal gland: 15.56% on BTCV, 29.01% on AbdomenCT), leading to unsatisfactory overall segmentation performance. Notably, compared to other methods, SAM2, without any fine-tuning, achieves comparable results by utilizing mask prompts and the memory attention mechanism, demonstrating the strength of SAM2's memory mechanism and laying a solid foundation for the success of our training-free method. Compared to other methods, RevSAM2 achieves significant improvements in mDSC on the BTCV and AbdomenCT datasets, with gains of 10.03% (69.86% vs. 59.83%) and 12.18% (82.46% vs. 70.28%), respectively.

Domain Adaptation The BTCV dataset includes annotations for 14 organs, which encompass all 13 organs labeled in the AbdomenCT dataset. We further demonstrate the strong robustness of our method when using support images from different datasets. In Table 3, we conduct the following experiments: for segmenting 13 organs in BTCV (excluding veins), we use images from AbdomenCT as sup-

port images; similarly, for segmenting 13 organs in AbdomenCT, we use images from BTCV as support images. We compare these results with the domain adaptation capabilities of other methods. We use the same three groups of support images as in Tables 1-2, and the average results are presented. As shown in Table 3, RevSAM2 still achieves the best performance in the domain adaptation experiments compared to other methods (69.03% vs. 59.52% of HQ-SAM on BTCV, 77.47% vs. 68.10% of CAT-SAM on AbdomenCT). Notably, the strong generalization ability of UniverSeg allows it to achieve better segmentation performance on BTCV when using AbdomenCT as support, compared to directly using BTCV. This result may indicate that AbdomenCT provides a broader or more consistent representation that enhances the model's ability to generalize across datasets. However, the overall segmentation performance remains somewhat unsatisfactory, potentially due to the loss of information caused by image compression.

Synapse-CT and CHAOS-MRI As shown in the Table 4, RevSAM2 achieves state-of-the-art mDSC performance on both Synapse-CT and CHAOS-MRI datasets, with improvements of 8.9% (87.42% vs. 78.52%) and 2.38% (85.28% vs. 82.90%). Especially for the spleen, it achieves improvements of 14.89% (94.02% vs. 79.13%) and 9.45% (85.82% vs. 76.37%) on Synapse-CT and CHAOS-MRI,

Forw Prop	Query Info	Rand Select	Rev Prop	Support image number 10	5	1
✓				64.93	61.00	40.62
✓	✓			64.68	58.69	38.51
✓	✓	✓		65.31	61.45	40.26
✓	✓		✓	69.86	67.38	51.57

Table 5. Ablation study on the BTCV dataset. Forw Prop: Forward Propagation; Query Info: Using adjacent query slice information; Rand Select: Randomly selecting query slices as the conditional slices; Rev Prop: Selecting the conditional slices via reverse prop.
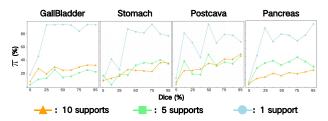

Figure 4. Line charts of π (%) versus the actual Dice (%) of p on the BTCV dataset when the number of supports is 10, 5, and 1.

respectively. It is worth mentioning that the support and query images for compared methods are paired one-to-one. For fair comparison, we stack all query images corresponding to the same support image to implement our method. More details are provided in the supplementary material.

4.4. Ablation Study

To verify the effectiveness of the two key steps in our method: (1) utilizing query volume's own information during inference (Query Self Propagation); and (2) selecting high quality query prediction via reverse propagation, we conduct the following ablation experiments, results are shown in Table 5.

Baseline Only the forward propagation stage in Sec 3.1 is used, and the dice score is calculated directly after obtaining P. Similar to most few-shot methods, this approach only uses the information from support images and ignores the latent information in the query volume.

Forward Prop. with Query Information In this experiment, support images and query volume are treated as conditional and non-conditional slices in Sec 3.3, respectively. The memory bank maintains a FIFO queue of query information to segment subsequent query slices, without distinguishing the quality of previous query prediction—whether good or bad. As table shows, this approach performs even worse than the baseline, as the FIFO queue gets contaminated with lower-quality predicted masks, which affects the segmentation of subsequent query images.

Random Selection of Query Information In this experiment, k query predictions are randomly selected to replace the top-k query predictions chosen through reverse propagation in Sec. 3.2. As shown in the table, the results are nearly identical to the baseline, indicating that random se-

	$k=9$	$k=7$	$k=3$	$k=1$
$N=10$	69.03	69.86	68.12	65.36
$N=5$	66.83	67.38	66.33	64.82
$N=1$	50.58	51.57	52.50	52.01

Table 6. Ablation of k on BTCV dataset. We conduct this ablation experiment under different support image numbers (N).

lection does not effectively choose high-quality query segmentation results.

Different Support Image Numbers As shown in the table, we conducted ablation study on support image numbers of 10, 5, and 1, and our method consistently achieved significant improvements in all cases. Figure 4 presents line charts showing the relationship between π values (explained in detail in Sec 3.2) and the actual Dice scores of p for four organs in BTCV. It can be observed that with 10 and 5 support images, the lines show a certain degree of positive correlation, but with 1 support image, the lines fluctuate significantly, and high π values appear even at low Dice scores, indicating that the top-k selections are more likely to include lower-quality p. This is because with only one support image in the memory bank, the memory attention mechanism is more susceptible to biases caused by the position and appearance of that support image. In other words, the information in the memory bank is less generalizable.

Number of k As described in Sec 3.3, we set the default value $k=7$. In Table 6, we conduct ablation experiments with $k=9$, $k=3$ and $k=1$ under different numbers of support images (N) to investigate the effect of k. As table shows, when $N=10$ or $N=5$, the differences between $k=7$ and $k=9$ is not sensitive, and as k decreases (at $k=3$ or $k=1$), the performance gradually declines. This observation aligns with the phenomenon observed in Fig 4: when $N=10$ or $N=5$, selecting high-quality p based on π is positively correlated, more conditional query slices helps improve query volume prediction. In contrast, when $N=1$, the p selected based on high π value may be of lower quality, and using more conditional query slices may hinder obtaining a good query volume prediction.

5. Conclusion

We propose RevSAM2, enabling SAM2 to perform medical image segmentation in data-scarce scenarios without any fine-tuning. By proposing reverse propagation, our RevSAM2 is able to select high-quality query predictions and employ these predictions as mask prompts to propagate within the query. Our RevSAM2 opens a new direction for leveraging SAM2. Notably, without any fine-tuning, RevSAM2 outperforms state-of-the-art few-shot algorithms under the one-shot setting, achieving superior segmentation performance. Its potential for application in data-scarce scenarios could provide an economical and efficient solution for medical image segmentation.

References

[1] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 6, 7

[2] Shurong Chai, Rahul Kumar Jain, Shiyu Teng, Jiaqing Liu, Yinhao Li, Tomoko Tateyama, and Yen-wei Chen. Ladder fine-tuning approach for sam integrating complementary network. *arXiv preprint arXiv:2306.12737*, 2023. 3

[3] Tianrun Chen, Ankang Lu, Lanyun Zhu, Chaotao Ding, Chunan Yu, Deyi Ji, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. *arXiv preprint arXiv:2408.04579*, 2024. 1, 3

[4] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*, 2023. 3

[5] Ziming Cheng, Shidong Wang, Tong Xin, Tao Zhou, Haofeng Zhang, and Ling Shao. Few-shot medical image segmentation via generating multiple representative descriptors. *IEEE Transactions on Medical Imaging*, 2024. 2, 6, 7

[6] Zhiheng Cheng, Qingyue Wei, Hongru Zhu, Yan Wang, Liangqiong Qu, Wei Shao, and Yuyin Zhou. Unleashing the potential of sam for medical adaptation via hierarchical decoding. In *CVPR*, 2024. 2, 3, 6, 7

[7] Guoyao Deng, Ke Zou, Kai Ren, Meng Wang, Xuedong Yuan, Sancong Ying, and Huazhu Fu. Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 368–377, 2023. 3

[8] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment anything model (sam) for digital pathology: Assess zeroshot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023. 3

[9] Hao Ding, Changchang Sun, Hao Tang, Dawen Cai, and Yan Yan. Few-shot medical image segmentation with cycleresemblance attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2488–2497, 2023. 3

[10] Ruiwei Feng, Xiangshang Zheng, Tianxiang Gao, Jintai Chen, Wenzhe Wang, Danny Z Chen, and Jian Wu. Interactive few-shot learning: Limited supervision, better medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(10):2575–2588, 2021. 3

[11] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P. Pereira, Matthew J. Clarkson, and Dean C. Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE Transactions on Medical Imaging*, 2018. 5

[12] Stine Hansen, Srishti Gautam, Robert Jenssen, and Michael Kampffmeyer. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 78:102385, 2022. 3

[13] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284, 2021. 2, 6, 7

[14] Chuanfei Hu, Tianyi Xia, Shenghong Ju, and Xinde Li. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation. *arXiv preprint arXiv:2304.08506*, 2023. 3

[15] 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

[16] 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*, 18(2):203–211, 2021. 2, 6, 7

[17] Kaidong Zhang, and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. 2, 3, 6, 7

[18] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 5

[19] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 6, 7

[20] 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. *arXiv:2304.02643*, 2023. 1, 3, 6

[21] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas 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. 5

[22] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas 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. 5

[23] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 5, 6

[24] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang,

Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 5

[25] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 3

[26] Jun Ma, Sumin Kim, Feifei Li, Mohammed Baharoon, Reza Asakereh, Hongwei Lyu, and Bo Wang. Segment anything in medical images and videos: Benchmark and deployment. *arXiv preprint arXiv:2408.03322*, 2024. 1, 3

[27] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(7):1837–1848, 2022. 3

[28] 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. *arXiv preprint arXiv:2408.00714*, 2024. 1, 6

[29] Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. 'squeeze & excite'guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020. 3

[30] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H Maier-Hein. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv preprint arXiv:2304.05396*, 2023. 3

[31] Qianqian Shen, Yanan Li, Jiyong Jin, and Bin Liu. Q-net: Query-informed few-shot medical image segmentation. In *Proceedings of SAI Intelligent Systems Conference*, pages 610–628, 2023. 3

[32] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 6

[33] Liyan Sun, Chenxin Li, Xinghao Ding, Yue Huang, Zhong Chen, Guisheng Wang, Yizhou Yu, and John Paisley. Few-shot medical image segmentation using a global correlation network with discriminative embedding. *Computers in biology and medicine*, 140:105067, 2022. 3

[34] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019. 3

[35] Runze Wang, Qin Zhou, and Guoyan Zheng. Few-shot medical image segmentation regularized with self-reference and contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 514–523, 2022. 2, 6, 7

[36] Huisi Wu, Fangyan Xiao, and Chongxin Liang. Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation. In *European Conference on Computer Vision*, pages 417–434, 2022. 2, 6, 7

[37] 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. *arXiv preprint arXiv:2304.12620*, 2023. 3

[38] Aoran Xiao, Weihao Xuan, Heli Qi, Yun Xing, Ruijie Ren, Xiaoqin Zhang, Ling Shao, and Shijian Lu. Cat-sam: Conditional tuning for few-shot adaptation of segment anything model. *arXiv preprint arXiv:2402.03631*, 2024. 2, 3, 6, 7

[39] Qinji Yu, Kang Dang, Nima Tajbakhsh, Demetri Terzopoulos, and Xiaowei Ding. A location-sensitive local prototype network for few-shot medical image segmentation. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 262–266, 2021. 3

[40] Yizhe Zhang, Tao Zhou, Shuo Wang, Peixian Liang, Yejia Zhang, and Danny Z Chen. Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–139, 2023. 3

[41] Jiayuan Zhu, Yunli Qi, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024. 1, 3, 6, 7

[42] Yazhou Zhu, Shidong Wang, Tong Xin, and Haofeng Zhang. Few-shot medical image segmentation via a region-enhanced prototypical transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 271–280, 2023. 2, 6, 7