

# SAM2-SGP: Enhancing SAM2 for Medical Image Segmentation via Support-Set Guided Prompting

Yang Xing  
*J. Crayton Pruitt Family Department of Biomedical Engineering University of Florida*

Jiong Wu  
*J. Crayton Pruitt Family Department of Biomedical Engineering University of Florida*

Yuheng Bu  
*Department of Electrical & Computer Engineering University of Florida*

Kuang Gong  
*J. Crayton Pruitt Family Department of Biomedical Engineering University of Florida*

**Abstract**—Although new vision foundation models such as Segment Anything Model 2 (SAM2) have significantly enhanced zero-shot image segmentation capabilities, reliance on human-provided prompts poses significant challenges in adapting SAM2 to medical image segmentation tasks. Moreover, SAM2’s performance in medical image segmentation was limited by the domain shift issue, since it was originally trained on natural images and videos. To address these challenges, we proposed SAM2 with support-set guided prompting (SAM2-SGP), a framework that eliminated the need for manual prompts. The proposed model leveraged the memory mechanism of SAM2 to generate pseudo-masks using image-mask pairs from a support set via a Pseudo-mask Generation (PMG) module. We further introduced a novel Pseudo-mask Attention (PMA) module, which used these pseudo-masks to automatically generate bounding boxes and enhance localized feature extraction by guiding attention to relevant areas. Furthermore, a low-rank adaptation (LoRA) strategy was adopted to mitigate the domain shift issue. The proposed framework was evaluated on both 2D and 3D datasets across multiple medical imaging modalities, including fundus photography, X-ray, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and ultrasound. The results demonstrated a significant performance improvement over state-of-the-art models, such as nnUNet and SwinUNet, as well as foundation models, such as SAM2 and MedSAM2, underscoring the effectiveness of the proposed approach. Our code is publicly available at [https://github.com/astlian9/SAM\\_Support](https://github.com/astlian9/SAM_Support).

**Index Terms**—Auto-prompting, Fine-tuning, Foundation Model, Medical Image Segmentation, SAM2.

## 1. Introduction

Vision foundation models have demonstrated strong zero-shot capabilities across various applications, including medical image segmentation [1]. Their impressive generalizability and few-shot learning capabilities make them attractive for adapting to downstream tasks, offering a more efficient alternative to training task-specific models

from scratch. The segment anything model (SAM) [2] is a recently developed visual foundation model designed for promptable image segmentation, pretrained on over 1 billion masks from 11 million natural images. Leveraging its large-scale training data and generalizable architecture, SAM exhibited strong zero-shot segmentation performance by using prompts as an extra input, such as a bounding box or positive and negative clicks, demonstrating exceptional generalization ability and establishing a new benchmark across various segmentation tasks [3], [4], [5]. Recent works also demonstrated the strong performance of the SAM model when applied to downstream medical image segmentation tasks [6], [7], [8], [9], [10], [11].

To extend these capabilities to more complex scenarios, the SAM2 model has been developed to expand the functionality of SAM to include video inputs [12]. This extension enabled SAM2 to process temporal sequences of images, making it suitable for tasks that required the understanding of spatial continuity over multiple frames. Fine-tuning it on specific tasks [13], [14], [15] and directly evaluating it on few-shot segmentation [16], [17], [18] are two ongoing research topics of SAM2 in medical image segmentation [19]. Although these SAM2-based methods required minimal or no training data, they still had notable limitations. Firstly, their performance remained highly dependent on user-provided high-quality instructions. Due to this limitation, recent work on the SAM2 model focused mainly on interactive medical image segmentation. Furthermore, it was trained on natural images and videos and could face domain shift issues when applied to medical image segmentation tasks.

To tackle the aforementioned limitations, we proposed a novel model, SAM2 with support-set guided prompting (SAM2-SGP), for medical image segmentation. By incorporating in-context learning with the memory mechanism of SAM2, SAM2-SGP could automatically generate high-quality prompts based on support sets. Specifically, the proposed SAM2-SGP model included a novel generator adapted from SAM2’s memory mechanism to generate the pseudo-mask from image-mask pairs of the support set. These pseudo-masks were then used to compute bounding boxes, which could be used to generate prompt embeddings.

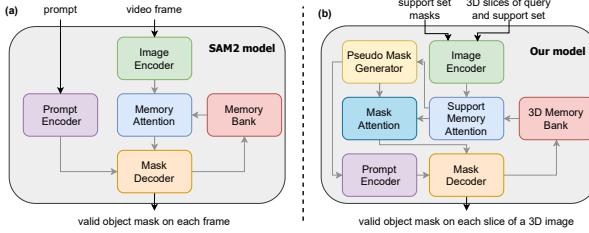


Figure 1. A comparison between (a) SAM2 and (b) the proposed SAM2-SGP. More details of SAM2-SGP are shown in Fig. 2.

Finally, image embeddings from the image encoder, pseudo-masks, and prompt embeddings from bounding boxes were subsequently processed by the pseudo-mask attention module for segmentation-map prediction.

Given the resemblance between video segmentation and 3D medical image segmentation, we also leveraged SAM2’s memory bank and memory attention modules to enable 3D image segmentation by treating 3D images as a temporal sequence of 2D images. A comparison between the proposed method and the original SAM2 model is shown in Fig. 1. Based on evaluations across multiple 2D and 3D medical imaging datasets, the proposed model consistently outperformed both fully supervised segmentation models and SAM2-based approaches. The major contributions of this work can be summarized as follows:

- We developed a prompt-free SAM2-based model for medical image segmentation. The proposed SAM2-SGP framework incorporated in-context learning with SAM2 and could achieve superior performance compared to other reference methods.
- A pseudo-mask generation module adapted from SAM2’s memory mechanism was introduced to generate pseudo-masks of query images based on the support set, enabling prompt generation without user interaction.
- A novel pseudo-mask attention module was introduced to generate the bounding-box prompts from pseudo-masks and improve localized feature extraction by using the pseudo-masks to guide attention to relevant areas.
- The proposed model was evaluated on 2D and 3D datasets from different modalities, including optical, fundus photography, X-ray, ultrasound, CT, MRI and PET.

## 2. Related works

### 2.1. SAM2 model

The SAM2 model inherited three core components from the original SAM architecture: an image encoder, a prompt encoder, and a mask decoder. The image encoder utilized a hierarchical Vision Transformer (ViT) backbone, Hiera [20], to extract multiscale feature embeddings from the input

image. The prompt encoder processed various forms of user input, such as positive and negative clicks, bounding boxes, or dense masks, into a format suitable for segmentation guidance. The mask decoder refined segmentation predictions by integrating image and prompt features through bidirectional Transformer blocks. Building on this foundation, SAM2 extended the original SAM model to video segmentation by introducing a memory mechanism that enhanced temporal consistency across sequential video frames. This memory mechanism consisted of three key components: a memory encoder, a memory bank, and a memory-attention module. Unlike SAM, which processed each image independently, SAM2 encoded features and predicted masks from previous frames using the memory encoder and stored them in the memory bank. During inference, the memory-attention module retrieved relevant embeddings from the memory bank and integrated them with the current frame’s features, refining segmentation predictions with temporal context. To support efficient memory usage, the memory encoder also downsampled predicted masks before storage. This structured memory-attention design enabled SAM2 to scale effectively while improving segmentation accuracy and robustness, particularly in challenging scenarios involving occlusion and object motion.

Recent studies have increasingly explored the application of SAM2 to medical image segmentation. FS-MedSAM2 [13] applied few-shot learning to evaluate SAM2 on the Synapse CT dataset [21]. FATE\_SAM [14] leveraged DINOv2 [22] to select the most similar samples for few-shot learning-based image segmentation. Similarly, Zhao et al. [15] adopted DINOv2 to construct support sets and evaluated SAM2 on the STARE dataset. In terms of fine-tuning, current efforts remained focused on prompt-based segmentation, which typically relied on high-quality human-provided prompts. Notable examples include MedSAM2 [16], SAM2\_med\_3D [17], RevSAM2 [23], and BioSAM2 [18]. In contrast to these approaches, the proposed SAM2-SGP model leveraged in-context learning in conjunction with SAM2’s memory mechanism to enable automatic prompting for medical image segmentation, reducing the reliance on manual input.

### 2.2. In-context learning

In-context learning (ICL) is a paradigm in which a model leverages a small set of examples (support set) provided during inference to make predictions on new inputs (query set). Unlike traditional fine-tuning, where models undergo explicit weight updates, ICL allows models to dynamically adapt by conditioning on context examples, making it particularly useful for few-shot learning scenarios. In medical image segmentation, ICL has been explored in Transformer-based architectures [24], where models were adapted to new tasks by conditioning on input queries and task-specific demonstrations. Universeg [25] trained a CNN with inputs from both the query set and the support set so that the support set could be seen as task-specific demonstrations for predicting the query set. ICL-SAM [26] followed Uni-

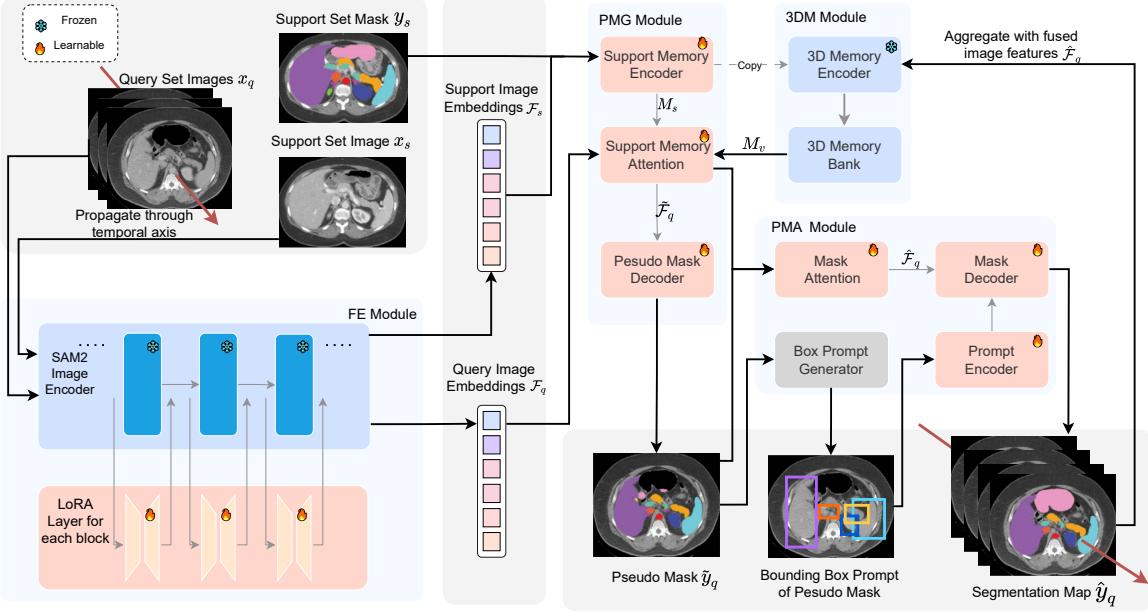


Figure 2. Diagram of the proposed framework. It contained four modules: (a) feature extraction (FE) module, (b) pseudo-mask generation (PMG) module, (c) pseudo-mask attention (PMA) module, and (d) 3D memory encoding (3DM) module. The 3D memory encoding module was only enabled when processing 3D datasets.

verseg's design and combined the ICL model and the SAM model for segmentation tasks. We observed that ICL conditioned query features on support set features, similar to how SAM2 conditioned current frame features on features from previous frames. Inspired by this connection, we proposed SAM2-SGP, which incorporated ICL into the SAM2 model. In SAM2-SGP, a subset of the training data was designated as the support set, which guided the generation of pseudo-masks used as the prompt.

### 3. Methodology

#### 3.1. Problem definition

Consider a segmentation task with a training set of  $N$  image-label pairs,  $S_{train} = \{(x_i^{train}, y_i^{train})\}_{i=1}^N$ , and a testing set of  $M$  image-label pairs,  $S_{test} = \{(x_j^{test}, y_j^{test})\}_{j=1}^M$ . A standard strategy in deep learning-based image segmentation is to learn a parametric function  $\hat{y} = f_\theta(x)$  to directly predict the segmentation map based on the input image  $x$ , where  $f_\theta(\cdot)$  indicates the network, and  $\theta$  denotes the network parameters. In our proposed framework, we randomly divided the training set into a support set  $S_{sup} = \{(x_k^{train}, y_k^{train})\}_{k=1}^D$  and a query set  $S_{qry} = S_{train} \setminus S_{sup}$ , where  $D = |S_{sup}|$  and  $|S_{qry}| \gg D$ . For each query image  $x_q \in S_{qry}$ , we selected a subset  $\tilde{S}_{sup} \subset S_{sup}$  consisting of the  $K$  most similar samples. Our goal was to train a function  $\hat{y}_q = f_\theta(x_q, \tilde{S}_{sup})$  that could accurately predict the segmentation map  $\hat{y}_q$  based on the input image  $x_q$  and its corresponding support subset  $\tilde{S}_{sup}$ .

For 3D image segmentation, we aimed to fully exploit spatial context by leveraging the memory mechanism of SAM2 to improve prediction quality and ensure inter-slice coherence. Given a 3D image  $x_q$  from  $S_{qry}$  with dimensions  $D \times H \times W$ , where  $D$  was the number of 2D slices and  $H \times W$  was the spatial size of each slice, the segmentation task involved training a parametric function  $\hat{y}_q^i = f_\theta(x_q^i, \hat{S}_{sup})$  to predict the segmentation map for slice  $i$ , where  $i \in 1, 2, \dots, D$ . Here, the dynamic support set  $\hat{S}_{sup}$  was defined as  $\hat{S}_{sup} = \tilde{S}_{sup} \cup \{(x_j^j, \hat{y}_q^j)\}_{j=1}^{i-1}$ , which combined: (1) the  $K$  most similar 2D slices from the static support subset  $\tilde{S}_{sup}$ , and (2) the previously seen slices of  $x_q$  along with their predicted segmentation maps. This formulation enabled the model to condition each slice's prediction not only on external support examples but also on the evolving context of the preceding slices within the same volume.

#### 3.2. Overall architecture

As shown in Fig. 2, the proposed model comprised four key components: the feature extraction (FE) module, the pseudo-mask generation (PMG) module, the 3D memory encoding (3DM) module, and the pseudo-mask attention (PMA) module. The feature extraction module processed images from both the query and support sets, extracting multiscale features. The pseudo-mask generation module encoded the extracted image embeddings of the support set with the corresponding masks to generate support memories. These support memories were then fused with the query

image embeddings and passed through a decoder to produce a pseudo-mask for the query image. The pseudo-mask attention module leveraged this pseudo-mask to generate a bounding box prompt, applied mask attention between the pseudo-mask and query embeddings, and outputted a segmentation map based on the box prompt, pseudo-mask and query embeddings. For the 3D memory encoding module, the prediction of the current slice was encoded to obtain the image embeddings and appended to the support memory to predict the pseudo-mask of the next slice. The details of each module are explained below.

### 3.3. Feature extracting module

The feature extraction module was supplied with images from both the query and support sets to extract multiscale features. The architecture was based on the SAM2 image encoder, which utilized Hiera [20] as the backbone. Given that the image encoder constituted approximately 70% of SAM2’s total parameters, we integrated Low-Rank Adaptation (LoRA) [27] layers to enable efficient fine-tuning. LoRA achieved this by decomposing weight updates into low-rank matrices, allowing adaptation with a reduced number of trainable parameters while preserving the pre-trained model’s knowledge. During the training process, the weight of the SAM2 image encoder remained frozen and only the LoRA layers were fine-tuned. The image embeddings of the query set  $\mathcal{F}_q$  and the support set  $\mathcal{F}_s$  were denoted as

$$\begin{aligned}\mathcal{F}_q &= \mathcal{E}(x_q), \\ \mathcal{F}_s &= \mathcal{E}(x_s),\end{aligned}\tag{1}$$

where  $\mathcal{E}(\cdot)$  was the SAM2 image encoder with LoRA layers.

### 3.4. Pseudo-mask generation module

SAM2’s memory components were originally developed to capture relationships between visual entities across images (i.e., tracking moving muscles in a ultrasound video), providing a robust mechanism for cross-image feature correlation. Inspired by in-context learning, this architectural capability was integrated into our proposed model to implement cross-image attention that enabled extraction of relevant query-image information conditioned on the corresponding features in the support set. To leverage this capability, the PMG module adapted SAM2 model’s memory encoder and memory attention blocks to generate initial pseudo-masks that guided subsequent processing stages. This module had three components: the support-memory encoder, the support-memory attention block, and the pseudo-mask decoder. The support-memory encoder, adapted from the SAM2 memory encoder, extracted memory features from the support set. It combined multiscale features of the support set,  $\mathcal{F}_s$ , with segmentation masks from the support set,  $y_s$ , through

$$\mathcal{M}_s = \phi(y_s) + \mathcal{F}_s,\tag{2}$$

where  $\phi(\cdot)$  was the convolutional module downsampling  $y_s$  to the same dimension as  $\mathcal{F}_s$ . The support-memory attention

block aligned the encoded support-set features,  $\mathcal{M}_s$ , with query-set features,  $\mathcal{F}_q$ , to produce query-specific feature representation  $\tilde{\mathcal{F}}_q$ , which could be denoted as

$$\begin{aligned}\tilde{\mathcal{F}}_q &= MA(\mathcal{M}_s, \mathcal{F}_q) = CA(\mathcal{M}_s, SA(\mathcal{F}_q)), \\ CA(x_1, x_2) &= softmax\left(\frac{Q_1 K^T}{\sqrt{d}} V_2 + V_2\right), \\ Q_1 &= W^Q x_1, \quad K_2 = W^K x_2, \quad V_2 = W^V x_2, \quad (3) \\ SA(x) &= softmax\left(\frac{Q K^T}{\sqrt{d}} V + V\right), \\ Q &= W^Q x, \quad K = W^K x, \quad V = W^V x,\end{aligned}$$

where  $MA(\cdot)$  was the memory attention module, which first applied a self-attention block with residual paths,  $SA(\cdot)$ , and then further went through a cross-attention block with residual paths,  $CA(\cdot)$ .  $W^Q, W^K, W^V$  denoted learnable projection matrices in the corresponding attention blocks that transformed input features into  $Q$  (query),  $K$  (key),  $V$  (value), respectively. The pseudo-mask decoder processed the encoded query set features  $\tilde{\mathcal{F}}_q$  to generate a pseudo-mask  $\tilde{y}_q$  for the query image using an empty prompt input, which could be written as

$$\tilde{y}_q = \mathcal{D}(\tilde{\mathcal{F}}_q, \mathcal{PE}(None)),\tag{4}$$

where  $\mathcal{D}(\cdot)$  was the lightweight decoder from SAM2 and  $\mathcal{PE}(\cdot)$  was the prompt encoder.

### 3.5. 3D memory module: an extra path to save information from previous slices

Processing 3D medical images in SAM2 was similar to processing video data, given the strong association between neighboring slices in 3D medical images. In our proposed network, the memory mechanism of the PMG module was extended to the 3D memory module. The 3D memory encoder encoded and stored features of previous slices and their corresponding predictions, which were stored in a 3D memory bank, denoted as

$$\mathcal{M}_v = \phi(\hat{y}_v) + \hat{\mathcal{F}}_v,\tag{5}$$

where  $\mathcal{M}_v$  represented the encoded volumetric memory from the predictions of adjacent slices  $\hat{y}_v$  and their corresponding image embeddings  $\hat{\mathcal{F}}_v$ . Such volumetric memory was then added to the support memory obtained from Eq (2) as

$$\mathcal{M}_s = [\mathcal{M}_s, \mathcal{M}_v].\tag{6}$$

The updated support memory,  $\mathcal{M}_s$ , was utilized in Eq (3) to generate the feature embeddings for the next slice. Through these modules, the predictions and the image embeddings of previous slices would be integrated into the support-set buffer to help compute the prediction of the current slice for 3D image segmentation. Notably, the memory bank size was set equal to the support-set size by default. During the training and inferring process, as the model propagated along the axial slices of the 3D image, the memory bank worked as a queue that accepted a new memory feature from

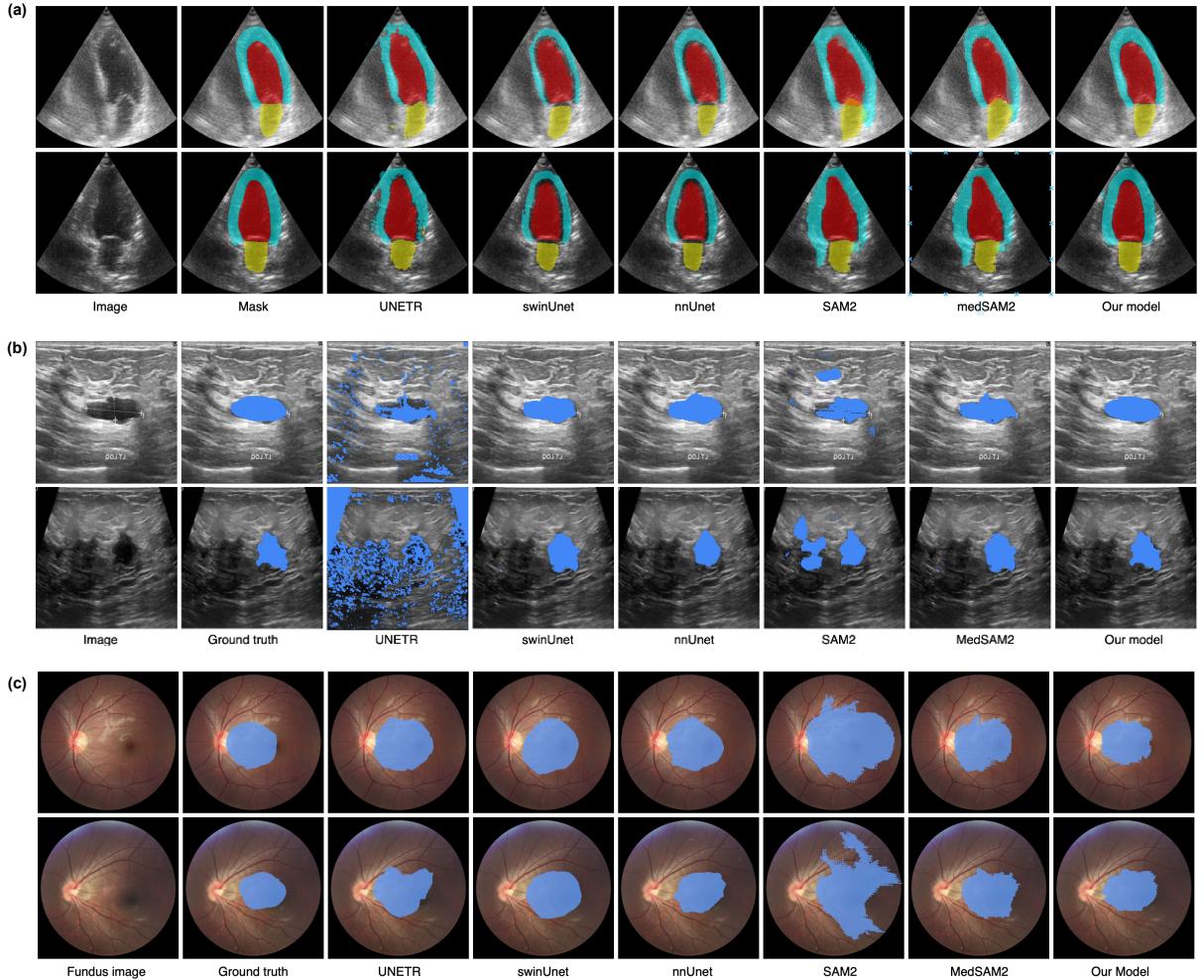


Figure 3. Examples of segmentation results from the CAMUS, BUSI, and REFUGE datasets. For each sub-figure, the first column shows the image, and subsequent columns present results from ground truth and 6 comparison methods. The two rows represent two different cases. (a) Segmentation results of the CAMUS dataset. The left ventricle, the atrium, and the myocardium were labelled in red, yellow, and cyan, respectively. (b) Segmentation results of the BUSI dataset. Tumors were segmented out in blue. (c) Segmentation results of the REFUGE dataset. Optic disc was labeled in blue.

the current axial slice and popped out an old memory feature. To improve the effectiveness of the memory-attention mechanism, we designed a memory bank that popped out the least similar memory based on the similarity score of all encoded memories and the image embeddings of the current slice.

### 3.6. Pseudo-mask attention module

The PMA module utilized pseudo-masks and image embeddings to generate the final prediction masks, with attention mechanisms playing a critical role. Transformer-based methods often exhibit slow convergence due to their reliance on global attention, which processes the entire image context indiscriminately. To address this limitation by leveraging the availability of the generated pseudo-masks, we proposed

a mask-attention module inspired by Mask3D [28] and Mask2Former [29]. This module constrained the attention mechanism to focus on the pseudo-mask regions, enabling more effective extraction of localized features rather than the entire global image context. Mask2Former employed a two-stage decoding process where the first stage produced a coarse mask and the second stage applied this coarse mask to guide attention operation using a mask-attention module before the fine-granularity pixel decoder. Our approach extended this concept by replacing the coarse mask with the pseudo-masks generated by the PMG module. Therefore, the computational complexity was significantly reduced by focusing operations on semantically relevant regions, which was particularly important for medical images where structures of interest often occupied a small portion of the overall image.

The PMA module consisted of three components: the SAM2 prompt encoder, the mask attention module, and the mask decoder. In the PMA module, the pseudo-masks were used for two purposes: (1) deriving bounding box prompts and (2) guiding the attention operation in the memory attention module following the Mask2Former’s design. A bounding box  $B_{\text{box}}$  was calculated based on the pseudo-mask  $\tilde{y}_q$ , and it was used to generate prompt embeddings via the SAM2 prompt encoder. As illustrated in Fig.2, the mask attention module fused the pseudo-mask with the image embeddings by leveraging the probabilistic map of the pseudo-mask to assign varying degrees of importance to different foreground regions. The mask attention module first resized the pseudo-mask  $\tilde{y}_q$  to the same dimension as the image embeddings  $\tilde{\mathcal{F}}_q$ , then applied self-attention on the image embeddings  $\tilde{\mathcal{F}}_q$ . An element-wise product was then performed between the resized pseudo-mask and the image embeddings, effectively suppressing irrelevant regions by multiplying a near-zero probability. The module was formulated as

$$\begin{aligned}\hat{\mathcal{F}}_q &= \tilde{y}'_q \odot \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V + \tilde{\mathcal{F}}_q, \\ Q &= W^Q \tilde{\mathcal{F}}_q, \quad K = W^K \tilde{\mathcal{F}}_q, \quad V = W^V \tilde{\mathcal{F}}_q\end{aligned}\quad (7)$$

where  $\tilde{y}'_q$  was the resized pseudo-mask, and  $W^Q, W^K, W^V$  represented the query, key, and value learnable projection matrices in the attention block, respectively.

As noted by MA-SAM [10] and H-SAM [30], the lightweight SAM decoder produced final predictions at a resolution 4 times lower than the input dimensions. The SAM2 model, employing an identical lightweight decoder structure, inherited the same limitation. To address this issue, we adopted a hierarchical pixel decoder following the design of H-SAM. Such decoder integrated the prompt embeddings from the prompt encoder and the outputs of the mask attention module to decode the image segmentation mask, denoted as:

$$\hat{y}_q = \mathcal{D}(\hat{\mathcal{F}}_q, \mathcal{PE}(B_{\text{box}})). \quad (8)$$

### 3.7. Loss function

The training loss consisted of the loss function between the final prediction and the ground truth, and the loss function between the pseudo-mask and the final prediction. Adding the second part was to make sure the pseudo-mask generation module generated pseudo-masks close enough to the final prediction. The loss function was formulated as

$$\mathcal{L} = \lambda_{dice} \mathcal{L}_{dice}(\hat{y}_q, y_q) + \lambda_{ce} \mathcal{L}_{ce}(\hat{y}_q, y_q) + \lambda_{KL} \mathcal{L}_{KL}(\tilde{y}_q, \hat{y}_q), \quad (9)$$

where  $\lambda_{dice}$ ,  $\lambda_{ce}$  and  $\lambda_{KL}$  were the weight hyper-parameters for each loss function,  $\mathcal{L}_{dice}$  denoted the dice loss function,  $\mathcal{L}_{ce}$  denoted the binary cross entropy loss function,  $\mathcal{L}_{KL}$  denoted the KL divergence loss function,  $\tilde{y}_q$  denoted the pseudo-mask from the PMG module,  $\hat{y}_q$  denoted the final prediction, and  $y_q$  denoted the ground truth.

| Model     | REFUGE       | PanDental    | WBC          | CAMUS        | BUSI         |
|-----------|--------------|--------------|--------------|--------------|--------------|
| UNETR     | 0.752        | 0.906        | 0.939        | 0.803        | 0.554        |
| SwinUnet  | 0.864        | 0.904        | 0.906        | 0.920        | 0.733        |
| nnUnet    | 0.849        | 0.929        | 0.951        | 0.925        | 0.751        |
| SAM2      | 0.753        | 0.854        | 0.627        | 0.830        | 0.577        |
| medSAM2   | 0.805        | 0.941        | 0.951        | 0.919        | 0.627        |
| Our model | <b>0.865</b> | <b>0.969</b> | <b>0.976</b> | <b>0.932</b> | <b>0.767</b> |

TABLE I. QUANTITATIVE RESULTS: DICE SCORES ON 2D DATASETS

## 4. Experiment

### 4.1. Datasets

Multiple public datasets were used to evaluate performance of the proposed model on both 2D and 3D tasks. The selected datasets covered a diverse range of medical imaging modalities and anatomical structures, with details below.

**4.1.1. 2D datasets.** 5 different 2D datasets were used, including: (1) White blood cell (WBC) Dataset [31] was a collection of microscopic images of white blood cells, used for segmentation and classification tasks. It contained 12,500 images with a resolution of 320×320 pixels, capturing various types of WBCs. (2) Panoramic dental dataset contained high-resolution panoramic dental radiographs annotated for tooth segmentation and bone segmentation [32]. It consisted of over 1,000 images with a resolution of approximately 2,000×1,000 pixels. (3) CAMUS dataset [33] was a cardiac ultrasound dataset with manual segmentation of the left ventricle, myocardium, and atrium. It contained 1,800 echocardiographic images with a resolution of 512×512 pixels. (4) BUSI dataset [34] was a breast ultrasound dataset containing images with tumor annotations for segmentation. It included 780 images with a resolution of 500×500 pixels. (5) REFUGE dataset [35] contained retinal fundus images designed for the segmentation of the optic disc and cup, which helped to detect glaucoma. The images had a resolution of 2,120×2,120 pixels.

**4.1.2. 3D datasets.** 3 different 3D datasets were utilized. (1) Head-and-neck PET/CT dataset [36] contained 100 3D PET images with a resolution of 124×124 pixels per slice, supporting the segmentation of tumors and organs at risk for radiation therapy planning. (2) The Automated Cardiac Diagnosis Challenge (ACDC) dataset [37] comprised cardiac MR images with segmentation annotations for the left and right ventricles and myocardium. It contained 150 cases with a resolution of 256×256 pixels for each slice. (3) Abdominal Multi-Organ Segmentation 2022 (AMOS22) dataset [38] was a multi-organ segmentation dataset based on abdominal CT scans, providing annotations for various organs. It included 500 cases with image resolutions ranging from 512×512 to 512×1024, covering 15 abdominal organs.

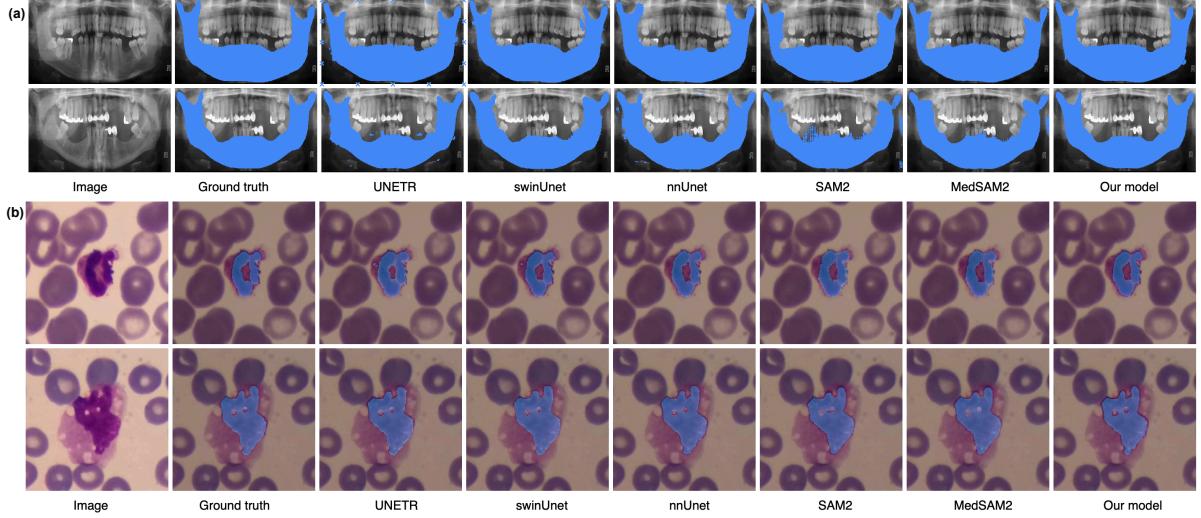


Figure 4. Examples of segmentation results from the Pandental and WBC datasets. For each sub-figure, the first column shows the image, and subsequent columns present results from ground truth and 6 comparison methods. The two rows represent two different cases. (a) Segmentation results of bones (labelled in blue) from the Pandental dataset. (b) Segmentation results of the nucleus (labelled in blue) from the WBC dataset.

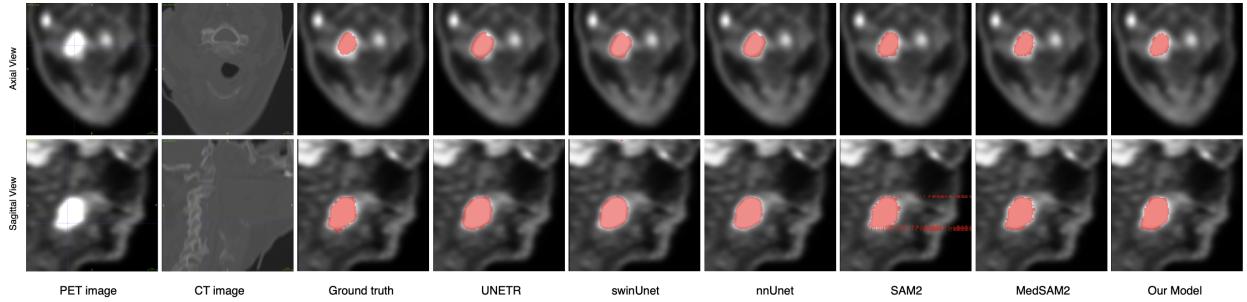


Figure 5. Examples of segmentation results from the PET/CT 3D dataset. The two rows represent the sagittal and axial views. The first column shows the PET image, the second column shows the CT image, and subsequent columns present segmentation results (labelled in pink) from ground truth and 6 comparison methods.

| Model     | AMOS_CT      | AMOS_MRI     | PET/CT       |
|-----------|--------------|--------------|--------------|
| UNETR     | 0.818        | 0.579        | 0.694        |
| SwinUnet  | 0.888        | 0.575        | 0.732        |
| nnUnet    | 0.863        | 0.676        | 0.724        |
| SAM2      | 0.699        | 0.560        | 0.682        |
| medSAM2   | 0.859        | 0.654        | 0.709        |
| Our model | <b>0.903</b> | <b>0.715</b> | <b>0.745</b> |

TABLE 2. QUANTITATIVE RESULTS: DICE SCORES ON THE AMOS22 AND PET/CT 3D DATASETS.

## 4.2. Evaluation and implementation details

Segmentation performance was evaluated using the Dice similarity coefficient (Dice). All models were trained using the Adam optimizer and NVIDIA A100 GPUs, with a single A100 GPU used for 2D datasets and four A100 GPUs for 3D datasets. Hyperparameters, including learning rates and batch sizes, were tuned for optimal performance

across different datasets. For 3D reference methods such as swinUnet and UNETR, which only accepted 3D images with a minimal patch size requirement of 32x32x32, each sample of the 2D dataset was repeated 32 times and stacked as a 3D sample as input.

## 5. Results

### 5.1. 2D segmentation results

Table 1 presents a quantitative comparison of the proposed model with several baseline methods, including UNETR [39], SwinUnet [40], nnUnet [41], SAM2 [12], and medSAM2 [16], on REFUGE, Panoramic Dental, WBC, CAMUS, and BUSI datasets. The proposed model consistently achieved superior performance across all evaluated 2D tasks. Note that for SAM-based methods, a random click

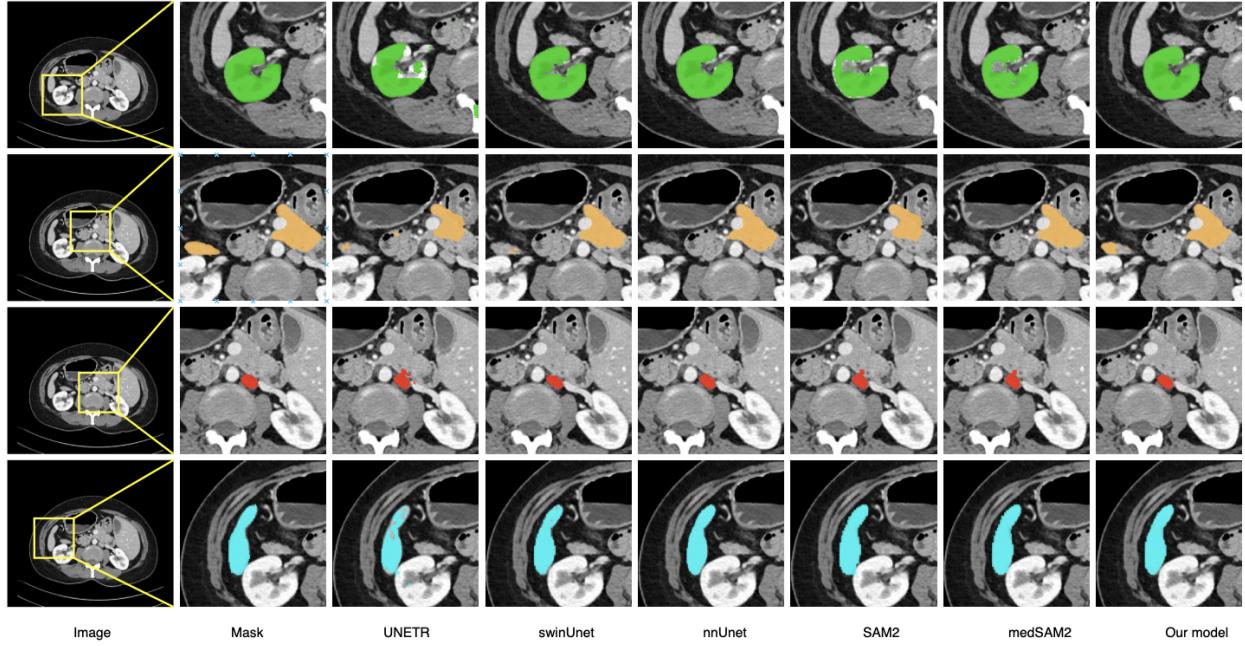


Figure 6. Examples of segmentation results from the AMOS22 3D dataset. Different organs were labeled in different colors, including left kidney (in green), spleen (in cyan), postcava (in orange), and pancreas (in red). The first column shows the image, and subsequent columns present results from ground truth and 6 comparison methods. The different rows came from the same CT slice.

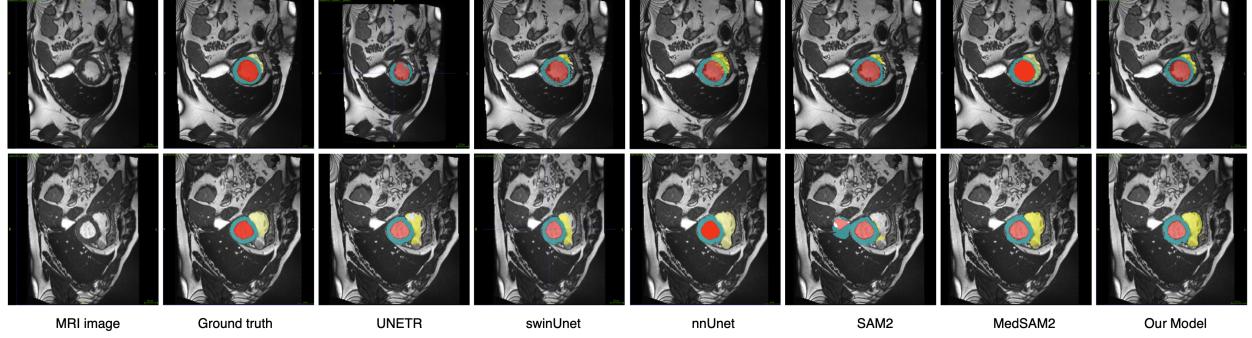


Figure 7. Examples of segmentation results from the ACDC 3D dataset. Left ventricle, right ventricle, and myocardium were labelled in red, yellow, and cyan, respectively. The first column shows the image, and subsequent columns present results from ground truth and 6 comparison methods. The two rows represent two different cases.

| Model     | Spleen       | Right Kidney | Left Kidney  | Gallbladder  | Esophagus    | Liver        | Stomach      | Aorta        | Post Cava    | Pancreas     |
|-----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| UNETR     | 0.927        | 0.885        | 0.901        | 0.665        | 0.733        | 0.941        | 0.787        | 0.914        | 0.682        | 0.745        |
| SwinUnet  | 0.955        | 0.938        | 0.945        | 0.773        | 0.831        | 0.960        | 0.889        | 0.947        | 0.749        | 0.849        |
| nnUnet    | 0.963        | <b>0.953</b> | 0.963        | 0.815        | <b>0.857</b> | 0.971        | <b>0.908</b> | 0.954        | <b>0.796</b> | <b>0.874</b> |
| SAM2      | 0.857        | 0.855        | 0.857        | 0.800        | 0.643        | 0.811        | 0.759        | 0.842        | 0.637        | 0.538        |
| medSAM2   | 0.956        | 0.940        | 0.941        | 0.792        | 0.725        | 0.947        | 0.758        | 0.831        | 0.772        | 0.798        |
| Our model | <b>0.966</b> | 0.952        | <b>0.969</b> | <b>0.882</b> | 0.852        | <b>0.974</b> | 0.888        | <b>0.962</b> | 0.793        | 0.851        |

TABLE 3. QUANTITATIVE RESULTS: DICE SCORES FOR 10 ORGANS ON THE AMOS22 3D DATASET.

| Model     | Average      | left heart   | right heart  | myocardium   |
|-----------|--------------|--------------|--------------|--------------|
| UNETR     | 0.866        | 0.940        | 0.853        | 0.865        |
| SwinUnet  | 0.900        | <b>0.958</b> | 0.885        | 0.856        |
| nnUnet    | 0.916        | 0.954        | 0.902        | <b>0.892</b> |
| SAM2      | 0.539        | 0.695        | 0.444        | 0.479        |
| MedSAM2   | 0.810        | 0.763        | 0.791        | 0.876        |
| Our model | <b>0.917</b> | 0.954        | <b>0.911</b> | 0.885        |

TABLE 4. QUANTITATIVE RESULTS: DICE SCORES ON THE ACDC 3D DATASET.

| Experiment | 3DM | PMG | PMA | left kidney  | right kidney |
|------------|-----|-----|-----|--------------|--------------|
| (a)        |     |     | ✓   | 0.936        | 0.928        |
| (b)        |     | ✓   | ✓   | 0.947        | 0.942        |
| (c)        | ✓   | ✓   | ✓   | <b>0.966</b> | <b>0.952</b> |

TABLE 5. ABLATION STUDY ON 3DM, PMG, PMA MODULES ON THE AMOS22 3D DATASET. PERFORMANCE WAS REPORTED IN DICE SCORES FOR LEFT AND RIGHT KIDNEYS.

was used as the prompt. Qualitative examples of the segmentation outcomes are illustrated in Fig.3 and Fig.4, further demonstrating the robustness and generalization capabilities of the proposed model in diverse 2D image segmentation scenarios.

## 5.2. 3D segmentation results

Table 2 summarizes the segmentation performance of various models on the PET/CT dataset [42], which shows that the proposed model achieved the best performance compared to other reference methods. Additionally, Table 2 compares Dice scores for CT and MRI images from the AMOS22 dataset, while Table 3 provides detailed organ-wise segmentation results specifically for CT images. The proposed model achieved the highest overall Dice scores on both CT and MRI images from the AMOS22 dataset, excelling in 5 out of 10 organs in the organ-wise segmentation evaluation, thus highlighting its effectiveness for multi-organ segmentation tasks. Table 4 presents Dice scores for the ACDC dataset, including both average and target-specific scores. Although the proposed model exhibited slightly lower performance in segmenting the left ventricle and myocardium, it achieved the highest overall Dice score, underscoring its competitive performance on 3D cardiac MRI segmentation. Qualitative segmentation examples are provided in Fig. 5 for the PET/CT dataset, Fig. 6 for the AMOS22 dataset, and Fig. 7 for the ACDC dataset. These qualitative examples align with the quantitative findings, further illustrating the model’s capability in generating accurate and consistent segmentation masks.

## 5.3. Ablation studies

We first performed ablation studies to evaluate the effectiveness of each component. Furthermore, additional ablation analyses were performed to investigate the impact of key fine-tuning settings, specifically: (1) support-set size, and (2) SAM2’s pretrained weights.

**5.3.1. Effectiveness of each component.** To evaluate the contribution of individual modules to the overall segmentation performance, an ablation study was conducted on the AMOS22 3D dataset, selectively disabling each component in turn. Table 5 presents the experimental results specifically for the left and right kidneys. Initially, both the 3DM and PMG modules, which were responsible for integrating contextual information from the support set and previous slices, were disabled, effectively reducing the model to a baseline SAM2 fine-tuned with a customized segmentation head. Experiment (a) revealed superior performance even at this baseline compared to advanced segmentation methods, highlighting the robustness of the SAM2 backbone. In experiment (b), only the 3DM module was disabled, and comparison to experiment (a) indicated notable improvements when the PMG module was enabled, validating the benefit of pseudo-mask guidance from the support set. Further comparison between experiments (b) and (c), where the 3DM module was subsequently enabled, demonstrated additional performance gains, emphasizing the importance of temporal memory from adjacent slices in the pseudo-mask generation process.

**5.3.2. Support-set size.** To examine how the support-set size affected performance, ablation experiments were designed by keeping the training set fixed and varying the support-set size. Specifically, a subset of training samples was duplicated to serve as the support set, ensuring that the training distribution remained unchanged. As shown in Table 6, enlarging the support set improved average Dice and Intersection over Union (IoU) scores, due to enhanced information provided by the larger support sets. This richer information resulted in more precise pseudo-mask generation through the support memory attention module. Nevertheless, increasing the support-set size also introduced additional computational overhead. Consequently, a support-set size of 4 was chosen as the default to achieve an optimal balance between performance and computational efficiency.

**5.3.3. SAM2 pretrained weights.** Ablation experiments were also conducted to evaluate the impact of initializing the model with the released weights of SAM2 versus training from scratch. Although SAM2 was originally trained on natural images, Table 7 demonstrates that the pretrained weights provided substantial benefits when transferred to medical domain, including better convergence and stronger zero-shot performance. These findings highlighted the strong generalizability and robustness of SAM2’s learned representations, even when applied to domain-shifted tasks such as medical image segmentation.

## 6. Discussion

In this work, we introduced a novel adaptation of SAM2 for medical image segmentation, leveraging the memory mechanism of SAM2 and in-context learning to address two major challenges of applying SAM2 model to medical image

| Support Size | Pandental     |               | CAMUS         |               | BUSI          |               | WBC           |               |
|--------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
|              | Dice          | IoU           | Dice          | IoU           | Dice          | IoU           | Dice          | IoU           |
| 1            | 0.9660        | 0.9370        | 0.9302        | 0.8710        | 0.7470        | 0.6706        | 0.9743        | 0.9546        |
| 2            | 0.9694        | 0.9406        | 0.9324        | 0.8751        | 0.7699        | 0.7010        | 0.9745        | 0.9552        |
| 4            | 0.9695        | 0.9409        | 0.9323        | 0.8749        | 0.7673        | 0.6925        | 0.9760        | 0.9579        |
| 8            | 0.9685        | 0.9391        | 0.9327        | 0.8755        | 0.7102        | 0.6340        | 0.9812        | 0.9634        |
| 16           | <b>0.9721</b> | <b>0.9457</b> | <b>0.9342</b> | <b>0.8777</b> | <b>0.7730</b> | <b>0.6924</b> | <b>0.9825</b> | <b>0.9659</b> |

TABLE 6. ABLATION STUDY ON DIFFERENT SUPPORT-SET SIZE. PERFORMANCE WAS REPORTED IN DICE SCORES AND IoU ACROSS DIFFERENT DATASETS.

| Dataset   | with SAM2 weights | without SAM2 weights |
|-----------|-------------------|----------------------|
| WBC       | 0.961             | 0.695                |
| Pandental | 0.951             | 0.801                |
| REFUGE    | 0.865             | 0.434                |
| BUSI      | 0.909             | 0.158                |
| CAMUS     | 0.930             | 0.427                |

TABLE 7. ABLATION STUDY ON THE EFFECT OF ADOPTING SAM2’s PRETRAINED WEIGHT. PERFORMANCE WAS REPORTED IN DICE SCORES ACROSS DIFFERENT DATASETS.

segmentation tasks: the reliance on high-quality human-guided prompts and the domain shift problem caused by SAM2’s training on natural images. By integrating a cross-attention module to generate pseudo-masks based on support sets, the proposed model automated the prompt generation and enhanced segmentation performance across various modalities, including fundus photography, X-ray, CT, MRI, PET, and ultrasound. The proposed method demonstrated significant improvements over other reference methods. Overall, the proposed approach represented a step toward automating and improving medical image segmentation using foundation models such as SAM2.

While the proposed method demonstrated promising performance, it also had certain limitations. First, the quality of the pseudo-masks was dependent on the quality of the support set used, which might affect the model’s robustness in scenarios with limited labeled data. Second, the strategy of applying the proposed model on the 3D dataset was to treat 3D images as videos, which could be considered as a temporal sequence of slices along one direction. However, this unidirectional propagation, from the first to the last slice, leveraged only half of the adjacent contextual information when predicting each slice, potentially limiting segmentation accuracy. Exploring strategies such as bidirectional propagation to process 3D datasets remains an area for future investigation. Additionally, further refining the pseudo-mask generation process and extending the model to efficiently process multi-modal imaging, e.g., utilizing both PET and CT images as inputs for the PET/CT 3D dataset, deserves further investigations.

## 7. Conclusion

This study proposed a novel foundation model for medical image segmentation built upon the SAM2 architecture. By incorporating in-context learning and pseudo-mask

driven prompting, the proposed model effectively addressed the challenges of adapting SAM2 to the medical domain and eliminated the need for human guidance. Evaluations across a diverse set of 2D and 3D datasets demonstrated consistent improvements of the proposed framework over both traditional and SAM-based segmentation methods.

## References

- [1] R. Bommasani, D. A. Hudson, and et al., “On the opportunities and risks of foundation models,” *CoRR*, vol. abs/2108.07258, 2021. [Online]. Available: <https://arxiv.org/abs/2108.07258>
- [2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment anything,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.02643>
- [3] L. Zhang, X. Deng, and Y. Lu, “Segment anything model (sam) for medical image segmentation: A preliminary review,” in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023, pp. 4187–4194.
- [4] C. Zhang, J. Cho, F. D. Puspitasari, S. Zheng, C. Li, Y. Qiao, T. Kang, X. Shan, C. Zhang, C. Qin, F. Rameau, L.-H. Lee, S.-H. Bae, and C. S. Hong, “A survey on segment anything model (sam): Vision foundation model meets prompt engineering,” 2024. [Online]. Available: <https://arxiv.org/abs/2306.06211>
- [5] T. Shaharabany, A. Dahan, R. Giryes, and L. Wolf, “Autosam: Adapting sam to medical images by overloading the prompt encoder,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.06370>
- [6] J. Ma, S. Kim, F. Li, M. Baharoon, R. Asakereh, H. Lyu, and B. Wang, “Segment anything in medical images and videos: Benchmark and deployment,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.03322>
- [7] J. Wu, W. Ji, Y. Liu, H. Fu, M. Xu, Y. Xu, and Y. Jin, “Medical sam adapter: Adapting segment anything model for medical image segmentation,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.12620>
- [8] Y. Yang, X. Wu, T. He, H. Zhao, and X. Liu, “Sam3d: Segment anything in 3d scenes,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.03908>
- [9] S. Gong, Y. Zhong, W. Ma, J. Li, Z. Wang, J. Zhang, P.-A. Heng, and Q. Dou, “3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation,” *Medical Image Analysis*, vol. 98, p. 103324, Dec. 2024. [Online]. Available: <http://dx.doi.org/10.1016/j.media.2024.103324>
- [10] C. Chen, J. Miao, D. Wu, A. Zhong, Z. Yan, S. Kim, J. Hu, Z. Liu, L. Sun, X. Li, T. Liu, P.-A. Heng, and Q. Li, “Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation,” *Medical Image Analysis*, vol. 98, p. 103310, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1361841524002354>
- [11] W. Lei, X. Wei, X. Zhang, K. Li, and S. Zhang, “Medlsam: Localize and segment anything model for 3d ct images,” 2024. [Online]. Available: <https://arxiv.org/abs/2306.14752>

- [12] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, “Sam 2: Segment anything in images and videos,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.00714>
- [13] Y. Bai, B. Yun, Z. Chen, Q. Yu, Y. Xia, and Y. Wang, “RevSAM2: Prompt sam2 for medical image segmentation via reverse-propagation without fine-tuning,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.04298>
- [14] X. He, Y. Hu, Z. Zhou, M. Jarraya, and F. Liu, “Few-shot adaptation of training-free foundation model for 3d medical image segmentation,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.09138>
- [15] L. Zhao, X. Chen, E. Z. Chen, Y. Liu, T. Chen, and S. Sun, “Retrieval-augmented few-shot medical image segmentation with foundation models,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.08813>
- [16] J. Zhu, A. Hamdi, Y. Qi, Y. Jin, and J. Wu, “Medical sam 2: Segment medical images as video via segment anything model 2,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.00874>
- [17] C. Shen, W. Li, Y. Shi, and X. Wang, “Interactive 3d medical image segmentation with sam 2,” 2025. [Online]. Available: <https://arxiv.org/abs/2408.02635>
- [18] Z. Yan, W. Sun, R. Zhou, Z. Yuan, K. Zhang, Y. Li, T. Liu, Q. Li, X. Li, L. He, and L. Sun, “Biomedical sam 2: Segment anything in biomedical images and videos,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.03286>
- [19] Y. He, P. Guo, Y. Tang, A. Myronenko, V. Nath, Z. Xu, D. Yang, C. Zhao, D. Xu, and W. Li, “A short review and evaluation of sam2’s performance in 3d ct image segmentation,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.11201>
- [20] C. Ryali, Y.-T. Hu, D. Bolya, C. Wei, H. Fan, P.-Y. Huang, V. Aggarwal, A. Chowdhury, O. Poursaeed, J. Hoffman, J. Malik, Y. Li, and C. Feichtenhofer, “Hiera: A hierarchical vision transformer without the bells-and-whistles,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.00989>
- [21] M. 2015, “Multi-atlas abdomen labeling challenge,” 2015. [Online]. Available: <https://www.synapse.org/>
- [22] M. Oquab, T. Dariseti, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “Dinov2: Learning robust visual features without supervision,” 2024. [Online]. Available: <https://arxiv.org/abs/2304.07193>
- [23] Y. Sun, J. Chen, S. Zhang, X. Zhang, Q. Chen, G. Zhang, E. Ding, J. Wang, and Z. Li, “Vrp-sam: Sam with visual reference prompt,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.17726>
- [24] X. Shi, D. Wei, Y. Zhang, D. Lu, M. Ning, J. Chen, K. Ma, and Y. Zheng, “Dense cross-query-and-support attention weighted mask aggregation for few-shot segmentation,” 2022. [Online]. Available: <https://arxiv.org/abs/2207.08549>
- [25] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca, “Universeg: Universal medical image segmentation,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.06131>
- [26] J. Hu, Y. Shang, Y. Yang, X. Guo, H. Peng, and T. Ma, “Icl-sam: Synergizing in-context learning model and sam in medical image segmentation,” in *Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning*, ser. Proceedings of Machine Learning Research, N. Burgos, C. Petitjean, M. Vakalopoulou, S. Christodoulidis, P. Coupe, H. Delingette, C. Lartizien, and D. Mateus, Eds., vol. 250. PMLR, 03–05 Jul 2024, pp. 641–656. [Online]. Available: <https://proceedings.mlr.press/v250/hu24a.html>
- [27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” 2021. [Online]. Available: <https://arxiv.org/abs/2106.09685>
- [28] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, “Mask3d: Mask transformer for 3d semantic instance segmentation,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8216–8223.
- [29] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, “Masked-attention mask transformer for universal image segmentation,” 2022.
- [30] Z. Cheng, Q. Wei, H. Zhu, Y. Wang, L. Qu, W. Shao, and Y. Zhou, “Unleashing the potential of sam for medical adaptation via hierarchical decoding,” 2024. [Online]. Available: <https://arxiv.org/abs/2403.18271>
- [31] Z. M. Kouzehkanan, S. Saghari, S. Tavakoli, A. Mahloojifar, M. S. Nosrati, S. Ghaffari, S. Setayeshi, E. Karami, P. Sasanpour, M. H. Nasirpour, H. Yousefi, S. Jangjoo, S. Rahimzadeh, H. Rabbani, P. Nematollahy, Z. Amini, S. Nasresfahani, and A. Sadeghipour, “A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm,” *Scientific Reports*, vol. 12, p. 1123, 2022. [Online]. Available: <https://doi.org/10.1038/s41598-021-04426-x>
- [32] A. H. Abdi, S. Kasaei, and M. Meh dizadeh, “Automatic segmentation of mandible in panoramic x-ray,” *Journal of Medical Imaging*, vol. 2, no. 4, p. 044003, October 2015. [Online]. Available: <https://doi.org/10.1117/1.JMI.2.4.044003>
- [33] S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P. M. Jodoin, T. Grenier, C. Lartizien, and O. Bernard, “Deep learning for segmentation using an open large-scale dataset in 2d echocardiography,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2198–2210, September 2019. [Online]. Available: <https://doi.org/10.1109/TMI.2019.2900516>
- [34] W. Al-Dhabayani, M. Gomaa, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” *Data in Brief*, vol. 28, p. 104863, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352340919312181>
- [35] J. I. Orlando, H. Fu, J. Barbosa Breda, K. van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, J. Lee, X. Li, P. Liu, S. Lu, B. Murugesan, V. Narango, S. S. R. Phaye, S. M. Shankaranarayana, A. Sikka, J. Son, A. van den Hengel, S. Wang, J. Wu, Z. Wu, G. Xu, Y. Xu, P. Yin, F. Li, X. Zhang, Y. Xu, and H. Bogunović, “Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs,” *Medical Image Analysis*, vol. 59, p. 101570, 2020.
- [36] G. Y. Li, J. Chen, S.-I. Jang, K. Gong, and Q. Li, “Swincross: Cross-modal swin transformer for head-and-neck tumor segmentation in pet/ct images,” *Medical Physics*, vol. 51, no. 3, pp. 2096–2107, 2024.
- [37] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. İlsgüm, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin, “Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?” *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.
- [38] Y. Ji, H. Bai, C. GE, J. Yang, Y. Zhu, R. Zhang, Z. Li, L. Zhanng, W. Ma, X. Wan, and P. Luo, “AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation,” in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [Online]. Available: <https://openreview.net/forum?id=Vk4-HUnkEk>
- [39] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, “Unetr: Transformers for 3d medical image segmentation,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 574–584.

- [40] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," 2022. [Online]. Available: <https://arxiv.org/abs/2201.01266>
- [41] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021. [Online]. Available: <https://doi.org/10.1038/s41592-020-01008-z>
- [42] B. Yu, S. Ozdemir, Y. Dong, W. Shao, K. Shi, and K. Gong, "Pet image denoising based on 3d denoising diffusion probabilistic model: Evaluations on total-body datasets," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 541–550.