OFL-SAM2: Prompt SAM2 with Online Few-shot Learner for Efficient Medical Image Segmentation

Meng Lan[1], **Lefei Zhang**[2], **Xiaomeng Li**[1*]

[1] Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology
[2]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
{eemenglan, eexmli}@ust.hk, zhanglefei@whu.edu.cn

Abstract

The Segment Anything Model 2 (SAM2) has demonstrated remarkable promptable visual segmentation capabilities in video data, showing potential for extension to medical image segmentation (MIS) tasks involving 3D volumes and temporally correlated 2D image sequences. However, adapting SAM2 to MIS presents several challenges, including the need for extensive annotated medical data for fine-tuning and high-quality manual prompts, which are both labor-intensive and require intervention from medical experts. To address these challenges, we introduce OFL-SAM2, a prompt-free SAM2 framework for label-efficient MIS. Our core idea is to leverage limited annotated samples to train a lightweight mapping network that captures medical knowledge and transforms generic image features into target features, thereby providing additional discriminative target representations for each frame and eliminating the need for manual prompts. Crucially, the mapping network supports online parameter update during inference, enhancing the model's generalization across test sequences. Technically, we introduce two key components: (1) an online few-shot learner that trains the mapping network to generate target features using limited data, and (2) an adaptive fusion module that dynamically integrates the target features with the memory-attention features generated by frozen SAM2, leading to accurate and robust target representation. Extensive experiments on three diverse MIS datasets demonstrate that OFL-SAM2 achieves state-of-the-art performance with limited training data.

Code — https://github.com/xmed-lab/OFL-SAM2

Introduction

Medical Image Segmentation (MIS) is an important step in medical image analysis, as it can assist in downstream applications such as disease diagnosis and monitoring of disease progression (Li et al. 2018a,b, 2020). Recently, the Segment Anything Model (SAM) (Huai et al. 2025; Zhang et al. 2024) has gained significant attention for its powerful segmentation abilities and prompt-based interactions. Nonetheless, SAM's zero-shot performance in MIS is suboptimal due to the domain gap between natural images and medical images.

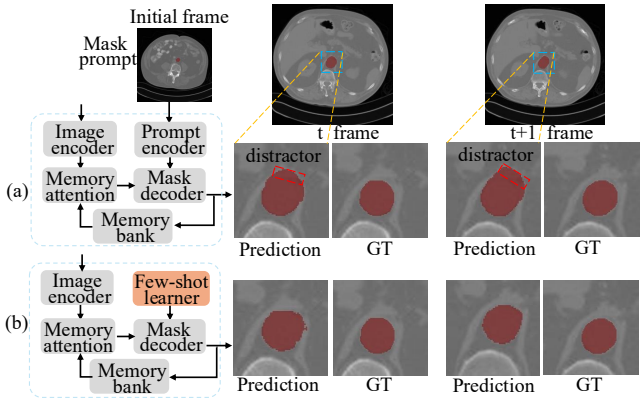

Figure 1: Comparisons of (a) SAM2 model and (b) our OFL-SAM2 model. SAM2 is susceptible to adjacent distractors.

To bridge this gap, previous works primarily focus on full fine-tuning (Ma et al. 2024; Zhu et al. 2024; Wang et al. 2025) or parameter-efficient fine-tuning (PEFT) (Xiao et al. 2024; Wang et al. 2024) using large annotated medical datasets. However, the fine-tuned medical SAM variants still require manual prompts for every frame, which requires sustained medical expert intervention, especially when processing 3D volumes as anatomically continuous 2D image sequences. SAM2 (Ravi et al. 2024) mitigates this issue through its streaming memory mechanism, enabling whole-sequence target segmentation with just a single-frame prompt, which has spurred interest in SAM2-based variants. Yet, these still rely on user-provided prompts, which hinders the automated processing of extensive images and is unfriendly to those without expertise. These limitations have driven the emergence of prompt-free SAM methods.

Current prompt-free medical SAM models typically replace manual prompts through three main strategies: (1) fine-tuning SAM models using PEFT strategies for direct semantic segmentation (Zhang and Liu 2023; Cheng et al. 2024), (2) employing prototype learning to acquire class-specific representative knowledge for self-prompt generation (Yue et al. 2024; Yan et al. 2025), (3) generating coarse masks via registration or regression methods from training sets, which are then converted into sparse/dense prompts (Xu et al. 2024). These prompt-free models process 3D vol-

*Corresponding author.

umes and temporally correlated 2D image sequences, e.g., surgical videos, in a frame-by-frame manner, providing a self-generated prompt for each frame, but they fail to leverage the spatiotemporal contextual information. In contrast, the SAM2 framework that effectively utilizes spatiotemporal features in sequential data shows promising research potential. One potential solution is to adapt existing prompt-free SAM methods to the SAM2 architecture. However, some of them, such as the leading H-SAM(Cheng et al. 2024), are incompatible with SAM2, as the memory encoder designed for binary mask prediction cannot encode their multi-class semantic predictions into the memory bank to guide the segmentation, whereas other methods are either limited by performance or require extensive fine-tuning data. Researchers have also explored the direct adaptation of SAM2 to 3D MIS (Bai et al. 2024b). For instance, the recently proposed FATE-SAM2 (He et al. 2025) retrieves images similar to the query image from the training set and combines them with past inference frames to construct memory embeddings. However, the performance of these prompt-free SAM2 variants remains unsatisfactory and is even inferior to leading prompt-free SAM-based methods. Upon careful investigation, we found that the core issue lies in the SAM2 framework itself, i.e., relying solely on the pixel-level feature matching based memory attention in SAM2 can make the model struggle to discriminate adjacent distractors, a critical challenge in medical images, where ambiguous boundaries frequently create distractors around targets, as shown in Fig. 1. This may be attributable to the fact that SAM2 is unable to provide discriminative target representations generated by a prompt for each frame as the SAM does.

In this study, we propose OFL-SAM2, an online prompt-free SAM2 framework for label-efficient MIS, aimed at providing discriminative target features for each frame like the prompt features of the SAM model by training a mapping network that can transform generic image features into target features using limited annotated samples, while preserving the temporal contextual information of SAM2. Based on SAM2, we introduces two lightweight yet effective components: 1) an online few-shot learner that trains the mapping network using limited data and updates the network parameters online during inference, and (2) an Adaptive Fusion Module (AFM) that dynamically integrates the target features with the memory-attention features to adapt the frozen decoder of SAM2 and suppress the potential distractor representations. Given the limited frame-mask pairs in the training set, OFL-SAM2 utilizes the image and memory encoders of SAM2 to extract the generic image features and the target features, which serve as the input and supervision for the mapping network, respectively. The mapping network is then optimized by the few-shot learner. During inference, OFL-SAM2 processes the query image by simultaneously generating target features through the mapping network and memory-attention features via SAM2, with AFM dynamically fusing these features for the decoder while employing a quality-aware selection mechanism to update both the mapping network and memory bank. Notably, the entire adaptation requires modest computational overhead as the mapping network and AFM each consist of just a single convolutional layer while keeping the core SAM2 components completely frozen, enabling rapid training convergence with limited annotations. Comprehensive evaluations across three diverse MIS datasets confirm that OFL-SAM2 achieves state-of-the-art performance for label-efficient MIS.

The primary contributions of this study are as follows:

- We propose a novel prompt-free SAM2 framework for label-efficient MIS, where we design an online few-shot learner that effectively utilizes both limited training samples and incoming test samples to continuously train a lightweight mapping network, enabling it to capture domain-specific medical knowledge and generate discriminative target representations for each frame in the sequence without requiring manual prompts.
- We develop an adaptive fusion module to dynamically integrate the learned target features with the memory-attention features to adapt the frozen SAM2 decoder and suppress the potential distractor representations.
- We evaluate OFL-SAM2 on three MIS datasets with multiple modalities, including CT, MRI, and surgical videos. The results demonstrate that OFL-SAM2 achieves state-of-the-art performance with limited training data.

Related Work

SAM-based Medical Image Segmentation

The SAM (Kirillov et al. 2023), trained on over a billion natural images, has demonstrated remarkable zero-shot segmentation capabilities when provided with visual prompts (e.g., a point or a bounding box). However, while SAM excels with natural images, its performance significantly degrades on medical images as evidenced by multiple studies (Deng et al. 2023; Cheng et al. 2023; Roy et al. 2023). This performance gap has spurred considerable research into adapting SAM for medical images, primarily through various fine-tuning approaches (Wu et al. 2025; Lin et al. 2024). For instance, MedSAM (Ma et al. 2024) created a large-scale medical image dataset to retrain SAM with bounding box prompts; SAMed (Zhang and Liu 2023) incorporated LoRA (Hu et al. 2022) layers into the image encoder while retaining the original mask decoder; H-SAM (Cheng et al. 2024) enhanced medical feature extraction through LoRA-modified encoders and introduced a hierarchical prompt-free decoder. MSA (Wu et al. 2025) developed specialized spatial and depth adapters for 3D medical images.

Building upon these SAM adaptations, researchers have similarly sought to optimize SAM2 for MIS. For example, MedicalSAM2 (Zhu et al. 2024) and MedSAM2 (Ma et al. 2025) fine-tuned the image encoder and mask decoder of SAM2 using extensive medical data. SAM2-Adapter (Chen et al. 2025) introduced lightweight adapters into the image encoder, which are fine-tuned together with the mask decoder. Some studies also explored the prompt-free SAM2 for MIS (Bai et al. 2024a). E.g., FATE-SAM2 (He et al. 2025) encoded some support image-mask pairs into the memory bank as memory features to launch the SAM2 inference directly. RevSAM2 (Bai et al. 2024b) proposed the reverse propagation strategy to select high-quality query information for the memory bank. In this work, we advance the

prompt-free adaptation of SAM2 for efficient MIS by introducing an online few-shot learner that generates discriminative target representations for each frame like a prompt while maintaining strong generalization across test sequences.

Online Discriminative Learning

The concept of online discriminative learning initially gained prominence in visual object tracking (Wang et al. 2021; Danelljan, Gool, and Timofte 2020), where its combination of strong performance and computational efficiency made it particularly effective, with these early approaches employing optimized convolutional filters trained online to perform robust foreground-background classification across video frames. This methodology was subsequently extended to video object segmentation, as demonstrated by FRTM (Robinson et al. 2020) which incorporated Conjugate Gradient and Gauss-Newton optimization to enable its few-shot learner to construct target-specific models from minimal template data during inference. The framework was further refined by LWL (Bhat et al. 2020a) through the introduction of a label encoder to generate information-rich few-shot labels, while JOINT (Mao et al. 2021) achieved complementary feature representation by combining transductive and online inductive features. More recently, this online learning paradigm has been successfully adapted to vision-language models, exemplified by Meta-Adapter (Song et al. 2023) which developed a lightweight residual-style adapter to enhance CLIP features under the guidance of a few-shot learner, showcasing the continued evolution and expanding applications of this approach across various domains.

Method

Overall Pipeline

Mathematically, we first divide a 3D volume into an anatomically continuous image sequence denoted as X, and its ith image is denoted as $x_i \in \mathbb{R}^{H\times W\times 3}$. Given the sequence of query images as $Q \in \mathbb{R}^{N\times H\times W\times 3}$, the training set $S = \{X_s, Y_s\}$, where $X_s = \{x_1, x_2, ...x_n\}$ represents the training images and $Y_s = \{y_1, y_2, ...y_n\}$ represents their corresponding labels, our goal is to predict the segmentation masks $P \in \mathbb{R}^{N\times H\times W}$ of Q using the limited training image-label pairs S. Notably, the training set may consist of multiple different medical image sequences.

As illustrated in Fig. 2, our OFL-SAM2 builds on SAM2 by removing the prompt encoder and including the designed few-shot learner and adaptive fusion module. OFL-SAM2 could be mainly divided into two branches: the online branch (the few-shot learner) and the offline branch (the memory attention module). In the training process, the training set is used to rapidly train the few-shot learner and the adaptive fusion module with the original modules of SAM2 frozen. During the inference, OFL-SAM2 sequentially processes the query images from sequence Q. The image encoder F_θ first extracts the generic features of the query image Q_i, which are sent to the two branches. The memory attention module produces the memory-conditioned feature $\mathbf{E}_1$, and the mapping network transforms the generic features into the target features $\mathbf{E}_2$. $\mathbf{E}_1$ and $\mathbf{E}_2$ are then intelligently fused through our adaptive fusion module to generate robust target features for the frozen decoder D_θ, yielding accurate final predictions. A quality assessment mechanism evaluates each output mask prediction to determine its suitability for updating both the memory bank and the mapping network parameters.

Memory Attention module

This offline branch directly utilizes the off-the-shelf memory attention module of SAM2, which performs the core cross-attention operation between the query image features and memory features in the memory bank to achieve the target information propagation. Since there is no user-provided prompt, we need to store some reference features in the memory bank to launch the memory attention module. To this end, we first select two image-mask pairs from the training set S that are most similar to the input query image Q_i. Given the image features $F_{Q_i} = F_\theta(Q_i)$ of Q_i and the image features $F_{S_j} = F_\theta(S_j)$ of the image x_j in training set S, we calculate the cosine similarity as follow:

$$\mathrm{Sim}\left(F_{Q_i}, F_{S_j}\right) = \frac{F_{Q_i} \cdot F_{S_j}}{\|F_{Q_i}\| \cdot \left\|F_{S_j}\right\|} \tag{1}$$

Based on the similarity scores, a ranked list of training image features is generated. Then the top K=2 most similar training image features and their corresponding masks are sent into the memory encoder to produce the memory features, which are then stored in the memory bank. These selected training examples may come from different training sequences, providing diverse reference anatomical information for the segmentation process.

The memory attention module contains four sequential memory attention layers, each layer comprises a self-attention module, a cross-attention module and a feed-forward network. In each attention layer, the query image features F_{Q_i} first go through the self-attention layer and then are fed into the cross-attention layer along with the memory features to retrieve and get the target representation from the memory features. Finally, the output features are obtained through the residual connection and feed-forward network. After four iterations of the memory attention layers, the memory-conditioned target features $\mathbf{E}_1$ are obtained. After the adaptive fusion module and the mask decoder, OFL-SAM2 generates the mask prediction of the query image.

Few-shot Learner

In this module, an online few-shot learner A_θ is designed to predict the parameters τ of the mapping network T_τ by minimizing squared error on the training samples, which can be formulated as follows:

$$L(\tau) = \frac{1}{2}\sum \|T_\tau\left(F_{S_i}\right) - M_{S_i}\|^2 + \frac{\lambda}{2}\|\tau\|^2. \tag{2}$$

Here, F_{S_i} is the generic features of the image x_i in the training set and M_{S_i} is the corresponding memory features. The mapping network T_τ, as a differentiable linear layer, where $\tau \in \mathbb{R}^{K\times K\times C\times D}$ is the weight of a convolutional layer with kernel size K, maps the C-dimension generic image feature to the D-dimension target-aware representation

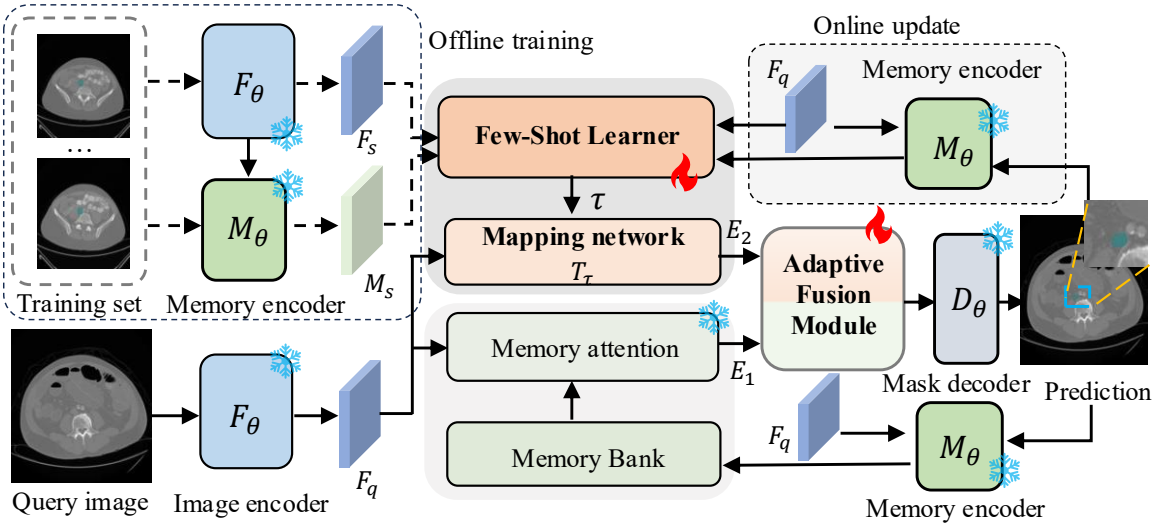

Figure 2: An overview of our OFL-SAM2 framework.

with the spatial size unchanged. $\frac{\lambda}{2}\|\tau\|^2$ is the regularization item, and λ is a learnable weight. The steepest descent method (Bhat et al. 2020b) is applied to iteratively minimize the squared error and optimize the weight τ. After the training of the few-shot learner, the trained mapping network is applied to the query image feature F_{Q_i} to generate the target representation $\mathbf{E}_2$. Before feeding into the subsequent adaptive fusion module, the channel dimension of $\mathbf{E}_2$ is converted from D to C by a convolutional layer to facilitate subsequent feature fusion. During inference, the high-quality mask prediction and its corresponding image features will be encoded as the training samples to update the mapping network parameters online through the few-shot learner.

Adaptive Fusion Module

Compared to the original SAM2, OFL-SAM2 introduces an additional feature flow generated by the mapping network. However, given that only a limited number of training samples are available, which are insufficient for effectively fine-tuning the decoder, we keep the decoder parameters frozen. In this case, how to effectively fuse the features of the two branches such that the data distribution of the fused features closely resembles that of the features input to the decoder in the original SAM2 is an urgent problem to be solved. At the same time, how to suppress the potential distractor representations in the features of the two branches, so that the fused features have a more robust and discriminative target representation is also a challenge to be considered. To address these issues, we propose the adaptive fusion module. Our AFM employs a weight network to learn an element-wise weight map W for the features of the offline branch, while simultaneously applying a complementary weight map $(1 - W)$ to the online branch features. This sophisticated weighting mechanism performs pixel-level feature recalibration, intelligently balancing the contributions from both branches and suppressing the irrelevant features to produce robust target features E_{tar} that maintain the original SAM2's expected data distribution. The AFM could be mathematically described as follows:

$$\begin{aligned} W &= G_\theta([E_1, E_2]), \\ E_{tar} &= W \odot E_1 + (1 - W) \odot E_2, \end{aligned} \tag{3}$$

where G_θ is the learnable weight network, $[\cdot]$ indicates the concatenation, and $\odot$ indicates element-wise multiplication. The weight network is implemented with a 3×3 convolution layer followed by a sigmoid function.

Update Strategy

Different from the strategy of SAM2 that indiscriminately stores all the predicted mask and the corresponding image features into the memory bank, we select high-quality predictions to update the memory bank and the mapping network parameters to prevent incorrect predictions from accumulating in the memory bank and misleading the mapping network. We adopt the confidence score to assess the quality of the mask prediction. Let $P_{(i,j)}$ denotes the probability of the mask prediction at location (i, j), and $\hat{G}$ represents the predicted binary mask, where $\hat{G}_{(i,j)}$ equals 1 when the probability of pixel at (i, j) is higher than a threshold of 0.5 and the pixel is predicted as the target object, otherwise it equals 0. we filter out the background and calculate the confidence score of the mask prediction as follows:

$$\mathcal{S}_{cf} = \frac{\sum_{i,j} P_{i,j} \cdot \hat{G}_{i,j}}{\sum_{i,j} \hat{G}_{i,j}} \tag{4}$$

If $\mathcal{S}_{cf}$ is higher than a threshold γ = 0.8, the prediction P and the corresponding query image features F_{Q_i} are fed into the memory encoder to produce the memory features M_{Q_i}. Then, F_{Q_i} and M_{Q_i} are sent into the few-shot learner as the training sample to update the mapping network parameters online, and M_{Q_i} is stored in the memory bank.

Method	Spleen	Right Kidney	Left Kidney	Gallb-ladder	Liver	Stomach	Aorta	Pancreas	Mean Dice↑(%)	HD ↓
SAM2 (Ravi et al. 2024)	78.24	82.67	77.94	66.42	75.59	72.35	68.76	48.31	71.29	24.08
Training set: 2 volumes										
SAMed (Zhang and Liu 2023)	84.62	80.70	81.57	62.60	90.44	66.33	77.50	51.46	74.41	21.49
SurgicalSAM (Yue et al. 2024)	87.03	85.31	86.72	70.44	85.18	64.77	85.46	47.98	76.61	17.42
H-SAM (Cheng et al. 2024)	88.28	82.79	84.00	69.95	**91.99**	**75.85**	83.68	55.71	79.03	15.32
FS-MedSAM2 (Bai et al. 2024a)	85.37	89.14	88.30	68.67	80.88	57.91	85.78	45.27	75.16	20.28
FATE-SAM2 (He et al. 2025)	85.49	89.59	88.61	68.59	81.60	58.55	86.01	45.28	75.47	19.10
OFL-SAM2	**90.11**	**90.43**	**89.62**	**72.64**	90.52	71.59	**88.45**	**61.17**	**81.81**	**9.10**
Training set: 3 volumes										
SAMed (Zhang and Liu 2023)	86.07	82.40	82.92	63.75	92.42	67.65	79.07	52.54	75.84	19.15
SurgicalSAM (Yue et al. 2024)	87.86	86.19	87.38	71.39	85.95	65.50	86.29	48.86	77.43	16.21
H-SAM (Cheng et al. 2024)	89.53	83.74	85.10	71.40	**93.59**	**76.65**	84.99	56.86	80.12	13.77
FS-MedSAM2 (Bai et al. 2024a)	86.11	89.91	89.12	69.80	81.72	58.53	86.63	45.88	75.96	18.90
FATE-SAM2 (He et al. 2025)	86.34	90.64	89.56	69.69	82.35	59.25	87.16	46.08	76.38	18.33
OFL-SAM2	**90.82**	**91.28**	**90.24**	**73.54**	91.27	72.19	**89.10**	**61.75**	**82.52**	**8.22**

Table 1: Comparison on Synapse-CT dataset. The Dice scores of the eight organs are reported. SAM2 adopts a mask prompt.

Experiments

Dataset and Evaluation

We conduct experiments on three medical datasets with different modalities, including Synapse-CT (Landman et al. 2015), PROMISE12 (Litjens et al. 2014) and Autolaparo (Wang et al. 2022). We utilize the Dice score (DSC) and the average Hausdorff distance (HD) as evaluation metrics.

The **Synapse-CT** multi-organ segmentation dataset has 30 cases in total and each CT volume contains 85 to 198 slices with a resolution of 512×512. Following SAMed (Zhang and Liu 2023), we evaluate eight abdominal organs. **PROMISE2012** dataset contains 50 3D transversal T2-weighted MR images of the prostate with manual binary prostate gland segmentation and is obtained from multiple centers with different acquisition protocols. The resolution of the PROMISE12 dataset is 512×512. **Autolaparo** is a surgical instrument segmentation dataset, which contains 300 laparoscopic video sequences and each sequence has 6 consecutive frames. Autolaparo focuses on 4 types of instruments and provides 8 types of segmentation annotations, where the shaft and manipulator of each instrument are annotated separately.

During the training process, for the Synapse-CT and PROMISE12 datasets, we construct two types of training sets by randomly selecting 3 and 2 volumes of data, and treat each 3D volume as an anatomically continuous 2D image sequence. For the Autolaparo dataset, we construct two types of training sets by randomly selecting 10% and 5% of the whole surgical video sequences and treating them as temporally correlated 2D image sequences. The remaining data of each dataset is used for evaluation.

Implementation details

We train our model on one NVIDIA RTX 3090 GPU using the SAM2_base model as our adapted SAM2 model. Considering the powerful segmentation capability of SAM2 and the limited labelled data, we chose to freeze the parameters of all SAM2 modules during the training process. The training loss is a combination of Cross-Entropy loss and Dice loss. The maximal training epoch is set to 40. The AdamW optimizer (Loshchilov and Hutter 2019) is employed for model optimization, with an initial learning rate of 1e-3. The learning rate decays by a factor of 0.1 at the 10th and 30th epochs. We conduct individual segmentation for each class in the datasets and report the average performance of three groups of random training data. In the few-show learner, C=256, D=64. Besides, during training, for each training sample, our few-shot learner employs 10 iterations of optimization to train the mapping network parameters T_τ with the default parameter initialization, while in the inference process, we reduce the number of iterations to 5.

Comparison with State-of-the-art Methods

To evaluate the performance of our proposed OFL-SAM2, we compared it with state-of-the-art methods on the three datasets. The comparison methods include the SAM2 model with mask prompt (Ravi et al. 2024), the prompt-free SAM variants: SAMed (Zhang and Liu 2023), SurgicalSAM (Yue et al. 2024), and H-SAM (Cheng et al. 2024), and the prompt-free SAM2 variants: FS-MedSAM2 (Bai et al. 2024a) and FATE-SAM2 (He et al. 2025).

Synapse-CT. As shown in Table 1, OFL-SAM2 exhibits outstanding performance on the Synapse-CT dataset under different training conditions. When using 3 training volumes, our OFL-SAM2 achieves 82.52% mean DSC and 8.22 HD, surpassing all competing methods by a signifi-

Training set	Method	Dice(%) ↑	HD ↓
None	SAM2 (mask)	80.44	15.26
2 volumes	SAMed	84.07	11.75
	SurgicalSAM	84.93	10.62
	H-SAM	85.44	9.30
	FS-MedSAM2	82.89	12.64
	FATE-SAM2	83.43	12.16
	OFL-SAM2	**88.07**	**7.18**
3 volumes	SAMed	86.12	10.57
	SurgicalSAM	86.74	9.61
	H-SAM	87.27	7.46
	FS-MedSAM2	83.88	12.03
	FATE-SAM2	84.58	11.18
	OFL-SAM2	**88.74**	**6.49**

Table 2: Comparison results on the PROMISE12 dataset.

Training set	Method	Mean Dice (%)↑	HD ↓
None	SAM2 (mask)	74.63	19.02
5%	SAMed	76.36	17.64
	H-SAM	80.22	14.48
	SurgicalSAM	80.76	13.32
	FS-MedSAM2	78.94	15.12
	FATE-SAM2	79.66	14.57
	OFL-SAM2	**83.78**	**11.88**
10%	SAMed	78.83	15.46
	H-SAM	82.80	12.76
	SurgicalSAM	83.04	12.33
	FS-MedSAM2	79.64	14.36
	FATE-SAM2	80.21	13.98
	OFL-SAM2	**84.41**	**11.04**

Table 3: Comparison results on the Autolaparo dataset.

cant margin, including a 6.14% DSC improvement over the SAM2 variant FATE-SAM2. When reducing the training set to merely 2 volumes, OFL-SAM2 maintains its performance advantage while exhibiting remarkable robustness, as evidenced by its minimal performance degradation of only 0.71% compared to H-SAM's 1.09% and SAMed's 1.43% drops, with this resilience likely stemming from OFL-SAM2's innovative online learning mechanism that enables dynamic adaptation to test sequences through continuous model refinement during inference. Some visualization comparisons of the multi-organ segmentation results are shown in the left part of Fig.3.

PROMISE12. Here, we compare the performance of our proposed OFL-SAM2 and the comparison methods on the PROMISE12 dataset. As presented in Table 2, without training, the SAM2 with a mask prompt attains 80.44% DSC. When trained on 2 and 3 volumes of data, our model achieves the DSC of 88.07% and 88.74%, respectively,

Dataset	Few-shot learner	AFM	Mean Dice↑ (%)
Synapse-CT	×	×	74.74
	✓	×	78.96
	✓	✓	**82.52**
PROMISE12	×	×	83.34
	✓	×	86.21
	✓	✓	**88.74**

Table 4: Ablation study on the few-shot learner module and AFM on the Synapse-CT and PROMISE12 datasets.

Dataset	Update strategy	γ	Mean Dice↑ (%)
Synapse-CT	×	-	81.48
	✓	0.7	82.13
	✓	0.8	**82.52**
	✓	0.9	82.04
PROMISE12	×	-	88.17
	✓	0.7	88.54
	✓	0.8	**88.74**
	✓	0.9	88.36

Table 5: Ablation study on the update strategy on the Synapse-CT and PROMISE12 datasets.

which outperforms the LoRA-fined SAM variants, such as H-SAM and SAMed, and the training-free FATE-SAM2. Under both the 2-volume and 3-volume training settings, OFL-SAM2 surpasses the second-place comparison method H-SAM by 2.63% and 1.47% in DSC, respectively, demonstrating the effectiveness of our model.

Autolaparo. We also evaluate the performance of our approach on the surgical video dataset, with comparison results presented in Table 3. When 10% of the video sequence data is used for training, OFL-SAM2 achieves state-of-the-art performance of 84.41% mean DSC and 11.04 HD, leading the second place SurgicalSAM by 1.37% and the SAM2 variant FATE-SAM2 by 4.2% on mean DSC. When only using 5% of data, OFL-SAM2 still achieves the best accuracy of 83.78% mean DSC, which is 3.56% higher than H-SAM. Some visualizations of the instrument segmentation results are shown in the right part of Fig.3.

Model Analysis

Module ablation. To verify the effectiveness of the proposed modules in our model, we perform ablation studies on the Synapse-CT and PROMISE12 datasets using 3 volumes of data. As shown in Table 4, the OFL-SAM2 without the few-shot learner and the AFM represents the model using only the memory attention module based on the most similar training image without training, which obtains 74.74% mean DSC on the Synapse-CT dataset. When we add the few-shot learner, the dynamically updatable mapping network provides the model with discriminative target representations and generalization to test sequences, which sig-

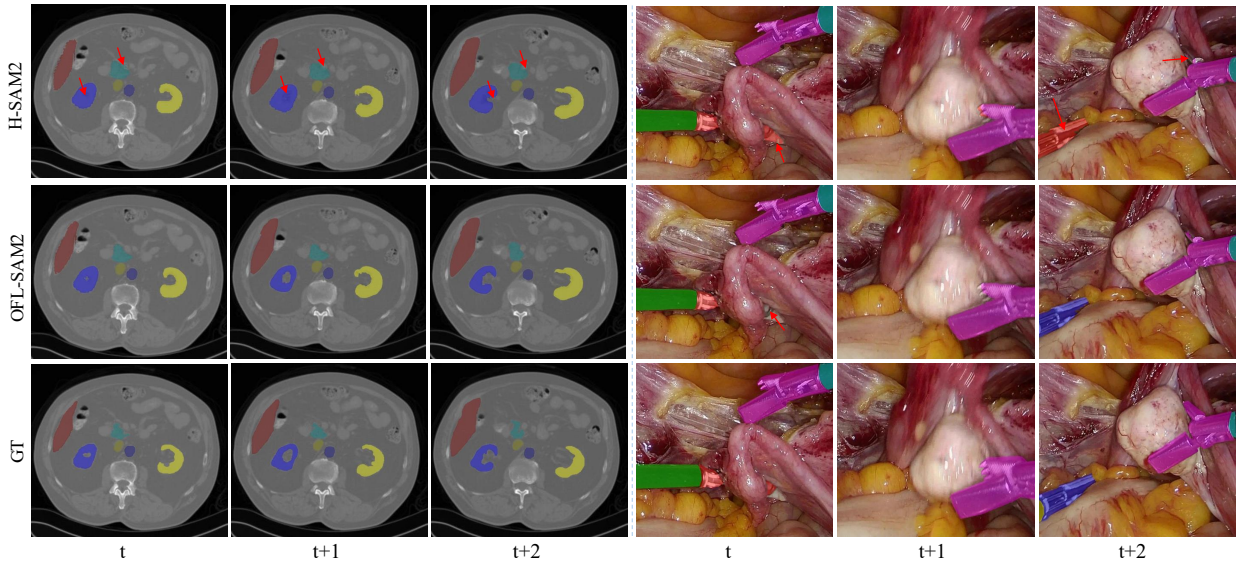

Figure 3: Visualization comparison between H-SAM and OFL-SAM2 on Synapse-CT (left) and Autolaparo (right) datasets.

Dataset	Image number (K)	Mean Dice↑ (%)
Synapse-CT	1	81.86
	2	**82.52**
	3	82.21
PROMISE12	1	88.27
	2	**88.74**
	3	88.68

Table 6: Ablation study on the number of most similar training images for memory bank.

nificantly improves the mean DSC by 4.22% to 78.96%. Furthermore, when both the few-shot learner and the AFM are included, the model can better fuse the features of the two branches to adapt the SAM2 decoder, thus realizing a better 82.52% mean DSC. These experiments prove the effectiveness of each module and also show that putting them together performs better.

Update strategy. Here, we explore the effectiveness of the proposed update strategy and the optimal hyperparameter γ selection within a certain range. We adopt the same data settings as module ablation and report the results in Table 5. It can be seen that there is a clear performance degradation of the model in the absence of the update strategy, whereas after using the strategy, the model achieves the optimal results on both datasets with $\gamma = 0.8$. When γ rises to 0.9, the accuracy drops. This might be that too high γ instead inhibits the selection of high-quality predictions, resulting in ineffective updates of the mapping network and memory bank.

Number of most similar training images. To investigate the influence of different numbers (K) of the most similar training images for the memory bank during inference, we conduct ablation experiments on the Synapse-CT and PROMISE12 datasets using 3 volumes of data, and the results are reported in Table 6. It can be seen that when K=2, OFL-SAM2 achieves the optimal performance. When K=1, the model's performance decreases by about 0.7% mean DSC on Synapse-CT and 0.5% DSC on PROMISE12. The reason may be that a single training image struggles to provide enough target information. When K=3, the performance stabilizes and declines slightly, possibly due to saturation of the provided target information and a reduction of feature information for subsequent updates. Therefore, we select K=2 as the default setting for better accuracy.

Conclusion

In this work, we introduce OFL-SAM2, a prompt-free SAM2 framework for efficient MIS of both 3D volumes and temporally correlated 2D image sequences. To achieve prompt-free operation while providing discriminative target representations akin to prompts for each frame, we propose an online few-shot learner that trains a lightweight mapping network to capture medical knowledge and transforms generic image features into target features using limited data. Critically, the few-shot learner can update the mapping network parameters during inference, enhancing the model's generalization across test sequences. Furthermore, to adapt the additional target representations to the frozen SAM2 framework and eliminate potential distraction representations, we devise an adaptive fusion module, which dynamically integrates the target features with the memory-attention features from SAM2, leading to accurate segmentations. Additionally, we propose a quality-aware update strategy that selectively incorporates high-confidence predictions for both the few-shot learner and memory bank updates to prevent error accumulation. Experimental results show that OFL-SAM2 achieves state-of-the-art performance on three diverse MIS datasets under data-limited conditions.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (NSFC) (Grant No. 62306254); the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China (Project No. R6005-24); and the Hong Kong Joint Research Scheme (JRS) of the NSFC/RGC (Project No. N_HKUST654/24).

References

Bai, Y.; Yu, Q.; Yun, B.; Jin, D.; Xia, Y.; and Wang, Y. 2024a. Fs-medsam2: Exploring the potential of sam2 for few-shot medical image segmentation without fine-tuning. *arXiv e-prints*, arXiv–2409.

Bai, Y.; Yun, B.; Chen, Z.; Yu, Q.; Xia, Y.; and Wang, Y. 2024b. RevSAM2: Prompt SAM2 for Medical Image Segmentation via Reverse-Propagation without Fine-tuning. *arXiv preprint arXiv:2409.04298*.

Bhat, G.; Lawin, F. J.; Danelljan, M.; Robinson, A.; Felsberg, M.; Van Gool, L.; and Timofte, R. 2020a. Learning what to learn for video object segmentation. In *ECCV*, 777–794.

Bhat, G.; Lawin, F. J.; Danelljan, M.; Robinson, A.; Felsberg, M.; Van Gool, L.; and Timofte, R. 2020b. Learning what to learn for video object segmentation. In *ECCV*, 777–794.

Chen, T.; Lu, A.; Zhu, L.; Ding, C.; Yu, C.; Ji, D.; Li, Z.; Sun, L.; Mao, P.; and Zang, Y. 2025. SAM2-Adapter: Evaluating & Adapting Segment Anything 2 in Downstream Tasks: Camouflage, Shadow, Medical Image Segmentation, and More. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Cheng, D.; Qin, Z.; Jiang, Z.; Zhang, S.; Lao, Q.; and Li, K. 2023. Sam on medical images: A comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035*.

Cheng, Z.; Wei, Q.; Zhu, H.; Wang, Y.; Qu, L.; Shao, W.; and Zhou, Y. 2024. Unleashing the potential of SAM for medical adaptation via hierarchical decoding. In *CVPR*, 3511–3522.

Danelljan, M.; Gool, L. V.; and Timofte, R. 2020. Probabilistic regression for visual tracking. In *CVPR*, 7183–7192.

Deng, G.; Zou, K.; Ren, K.; Wang, M.; Yuan, X.; Ying, S.; and Fu, H. 2023. Sam-u: Multi-box prompts triggered uncertainty estimation for reliable sam in medical image. In *MICCAI*, 368–377.

He, X.; Hu, Y.; Zhou, Z.; Jarraya, M.; and Liu, F. 2025. Few-Shot Adaptation of Training-Free Foundation Model for 3D Medical Image Segmentation. *arXiv preprint arXiv:2501.09138*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huai, Z.; Tang, H.; Li, Y.; Chen, Z.; and Li, X. 2025. Leveraging Segment Anything Model for Source-Free Domain Adaptation via Dual Feature Guided Auto-Prompting. *IEEE Transactions on Medical Imaging*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*, 4015–4026.

Landman, B.; Xu, Z.; Igelsias, J.; Styner, M.; Langerak, T.; and Klein, A. 2015. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, 12.

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

Li, X.; Yu, L.; Chen, H.; Fu, C. W.; and Heng, P. A. 2018b. Semi-supervised Skin Lesion Segmentation via Transformation Consistent Self-ensembling Model. In *British Machine Vision Conference (BMVC)*.

Li, X.; Yu, L.; Chen, H.; Fu, C.-W.; Xing, L.; and Heng, P.-A. 2020. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE transactions on neural networks and learning systems*, 32(2): 523–534.

Lin, X.; Xiang, Y.; Yu, L.; and Yan, Z. 2024. Beyond adapting SAM: Towards end-to-end ultrasound image segmentation via auto prompting. In *MICCAI*, 24–34.

Litjens, G.; Toth, R.; Van De Ven, W.; Hoeks, C.; Kerkstra, S.; Van Ginneken, B.; Vincent, G.; Guillard, G.; Birbeck, N.; Zhang, J.; et al. 2014. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical Image Analysis*, 18(2): 359–373.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Ma, J.; Yang, Z.; Kim, S.; Chen, B.; Baharoon, M.; Fallahpour, A.; Asakereh, R.; Lyu, H.; and Wang, B. 2025. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*.

Mao, Y.; Wang, N.; Zhou, W.; and Li, H. 2021. Joint inductive and transductive learning for video object segmentation. In *ICCV*, 9670–9679.

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

Robinson, A.; Lawin, F. J.; Danelljan, M.; Khan, F. S.; and Felsberg, M. 2020. Learning fast and robust target models for video object segmentation. In *CVPR*, 7406–7415.

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

Song, L.; Xue, R.; Wang, H.; Sun, H.; Ge, Y.; Shan, Y.; et al. 2023. Meta-adapter: An online few-shot learner for vision-language model. In *NeurIPS*, volume 36, 55361–55374.

Wang, H.; Guo, S.; Ye, J.; Deng, Z.; Cheng, J.; Li, T.; Chen, J.; Su, Y.; Huang, Z.; Shen, Y.; Fu, B.; Zhang, S.; He, J.; and Qiao, Y. 2025. SAM-Med3D: Towards General-Purpose Segmentation Models for Volumetric Medical Images. In *ECCV 2024 Workshops*, 51–67.

Wang, H.; Lin, Y.; Ding, X.; and Li, X. 2024. Tri-plane mamba: Efficiently adapting segment anything model for 3d medical images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 636–646. Springer.

Wang, N.; Zhou, W.; Wang, J.; and Li, H. 2021. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, 1571–1580.

Wang, Z.; Lu, B.; Long, Y.; Zhong, F.; Cheung, T.-H.; Dou, Q.; and Liu, Y. 2022. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *MICCAI*, 486–496.

Wu, J.; Wang, Z.; Hong, M.; Ji, W.; Fu, H.; Xu, Y.; Xu, M.; and Jin, Y. 2025. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102: 103547.

Xiao, A.; Xuan, W.; Qi, H.; Xing, Y.; Ren, R.; Zhang, X.; Shao, L.; and Lu, S. 2024. Cat-sam: Conditional tuning for few-shot adaptation of segment anything model. In *ECCV*, 189–206.

Xu, J.; LiXiaokang; Chengyuyue; Ma, C.; Guo, Y.; and Wang, Y. 2024. SAM-MPA: Applying SAM to Few-shot Medical Image Segmentation using Mask Propagation and Auto-prompting. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*.

Yan, Z.; Yin, Z.; Lin, T.; Zeng, X.; Liang, K.; and Ma, Z. 2025. PGP-SAM: Prototype-Guided Prompt Learning for Efficient Few-Shot Medical Image Segmentation. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, 1–5.

Yue, W.; Zhang, J.; Hu, K.; Xia, Y.; Luo, J.; and Wang, Z. 2024. Surgicalsam: Efficient class promptable surgical instrument segmentation. In *AAAI*, volume 38, 6890–6898.

Zhang, K.; and Liu, D. 2023. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*.

Zhang, Q.; Li, Y.; Xue, C.; Wang, H.; and Li, X. 2024. GlandSAM: injecting morphology knowledge into segment anything model for label-free gland segmentation. *IEEE Transactions on Medical Imaging*.

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