# MEI: Mutual-Enhanced Integration Between Pretrained SAM and Lightweight Model for Medical Image Segmentation


1st Zhiyan Wang
*College of Computer Science and Technology*
*National University of Defense Technology*
Changsha, China
wangzhiyan24@nudt.edu.cn

2nd Changjian Wang
*College of Computer Science and Technology*
*National University of Defense Technology*
Changsha, China
wangcj@nudt.edu.cn

3rd Zhongshun Tang
*Zhujiang Hospital of Southern Medical University*
Guangzhou, China
tangzhongshun2000@163.com

4th Maolin Luo
*College of Computer Science and Technology*
*National University of Defense Technology*
Changsha, China
luomaolin21@nudt.edu.cn

5th Daozhong Lei
*Hunan College of Information*
Changsha, China
leidaozhong@mail.hniu.cn

6th Qian Deng
*the 921st Hospital of the People's Liberation Army*
Changsha, China
csuchase@163.com



***Abstract*—The success of pretrained large models has provided new and important technical approaches for medical image segmentation (MedISeg). However, pretrained models also have two significant shortcomings: (i) high-cost fine-tuning is usually required when they are applied to specific tasks; (ii) it is difficult to incorporate new attributes not used during the pretraining process. To address these issues, we propose the Mutual-Enhanced Integration framework Between Pretrained SAM and Lightweight Model for MedISeg, named MEI. Firstly, Adaptive Prompt Enhancement (APE) module is proposed to optimize SAM-Med by enhancing prompt quality based on the inconsistency between SAM-Med and the lightweight model, thereby introducing new features to the pretrained SAM and improving its specificity on specific task datasets. Secondly, Ensemble Structure Boundary Enhancement (ESBE) module is designed to optimize the lightweight model by providing more accurate boundary information via the fusion of prediction results from both the pretrained SAM and the lightweight model, thereby enhancing the segmentation accuracy of the lightweight model. Extensive experiments on four public 2D/3D segmentation datasets demonstrate that our model achieves the state-of-the-art performance relative to a strong baseline.** [1]



***Index Terms*—Medical Image Segmentation, Pretrained SAM, Lightweight Model.**


## I. Introduction

Medical image segmentation (MedISeg) represents a fundamental yet highly challenging task [1], where deep learning has become the mainstream method. With the success of large model technology, the use of pre-training strategy to improve MedISeg has been paid more attention due to the demonstrated potential in the generalization ability.

SAM-Med [2] was a salient marker model in MedISeg which is proposed based on Segment Anything Model (SAM) [3]. It has shown promising generalization ability by introducing a prompt mechanism and powerful pre-training strategy. However, SAM-Med still present performance constraints similar to other pretrained models. The first, SAM-Med is usually not directly applied to specific tasks and fine-tuning strategy is effective while it can not form an absolute advantage over traditional lightweight models on specific task datasets. The second, SAM-Med is usually used in the embedding stage of medical images while its fixed pre-training mode will make it difficult to introduce new features into the embedding of medical images.

On the contrary, traditional lightweight models demonstrated robust task-specific adaptability on dedicated scenario datasets, although its weak generalization ability. They remain prevalent in current applications due to their low resource requirements and high computational speed. [4], [5]

Let's ponder the question: **Can we leverage the advantages of the lightweight model to improve the performance of SAM-Med?** Specifically, *Can new features be effectively introduced into SAM-Med through integrating the lightweight model for better medical image segmentation performance?*

Some studies have recognized the potential of traditional lightweight models to enhance pre-trained large models. Two methods have been proposed to improve SAM-Med: one [6]


[1]This work is funded with the Research Project BHQ090003000X03

uses semi-supervised learning and adds pseudo-labels to selected samples for training, and the other [7] integrates a lightweight model with SAM and fuses their results. However, both them also have significant limitations and cannot effectively introduce new features. Additionally, some researchers have employed the pre-trained SAM to enhance lightweight models. One approach [8] involves using SAM to generate boundary map for augmenting the training images fed into the model. However, this may fail to improve performance because the boundary features generated by SAM might be incompatible with the lightweight model's learning capacity.

In this paper, we propose the Mutual-Enhanced Integration framework Between the Pretrained SAM and the Lightweight MedISeg Model, named MEI, which can effectively leverage complementarity between SAM-Med and a traditional lightweight segmentation model and integrate the inherent features in medical images: Firstly, APE module is proposed to optimize SAM-Med's prompt information according to the discrepancy between the output of the lightweight backbone and SAM-Med inference results. Additionally, the intermittently selected lightweight backbone's high-quality results are fed into the prompt encoder of SAM-Med. Secondly, ESBE mechanism is proposed to optimize the lightweight segmentation model according to SAM-Med's output. In the meanwhile, a novel adaptive integration attention (AIA) module is designed for the enhancement to fuse boundary information from lightweight backbone's segmentation result and SAM-Med's inference results to reduce pixel level errors in MedISeg.

To summarize, our contributions are:

- An APE mechanism is proposed for SAM-Med enhancement which can dynamically determine the prompt by comparing inconsistencies between our lightweight backbone's outputs and the SAM-Med's inference results. It also intermittently incorporates optimal prompts to ensure the accuracy of the SAM-Med's prompt. This mechanism can enhance the specificity of SAM-Med.
- An ESBE mechanism is proposed to generate refined boundaries based on the new designed AIA module. It can effectively fuse boundary results from both our lightweight backbone and SAM-Med, and then enhances the lightweight model to achieve higher generality.
- Extensive evaluations on four diverse public 2D/3D segmentation datasets demonstrate that our model achieves a series of state-of-the-art segmentation outcomes. Furthermore, comprehensive ablation studies validate the effectiveness of each proposed module.

**The source code will be available after acceptance.**

## II. Related Work

### A. Traditional and Deep-learning based Methods

Traditional MedISeg methods primarily exploit the inherent dissimilarities between foreground and background regions to guide the segmentation. According to the different features used, they can be divided into several categories such as OTSU, region growing algorithms, level set algorithms [9], co-occurrence matrix algorithms [10], gray texture matrix algorithms [11], etc. Although they are no longer research hotspots, some studies have still made progress in recent years. [12] developed an enhanced 2D OTSU algorithm incorporating modified particle swarm optimization (PSO). [13] proposed an improved region growing approach which can constrain the growth region by using prior knowledge with automatic threshold detection, thereby optimizing the segmentation outcomes. [14] introduced an advanced variational level set method which employs a selective binarized Gaussian filter as the pressure function. However, traditional methods use artificial features which are limited by our cognition, resulting in frequently inadequate segmentation precision and applicability.

However, current deep learning-based approaches primarily focus on utilizing multilayer networks to extract high-level feature representations. Notably, specialized networks derived from the UNet focus on critical segmentation regions. These developments have led to classical variants including FCN, VNet, UNet++ [15], and Attention-UNet [16], as well as more advanced models incorporating the Transformer such as Swin-UNet [17], UNETR [18], UNETR++ [19], and Swin-UNETR [20], with their core innovations focusing on optimizing the feature extraction process [21]–[23] and skip-connection [24], [25]. Since the single-model methods often suffer from limited accuracy, the strategy [29]–[31] combining multiple models is proposed to compensate for limitations in feature extraction capability.

### B. Pretrained Foundation Model-based Methods

The introduction of the Segment Anything Model (SAM) [3] spurred the development of numerous SAM-based foundation models dedicated to medical image segmentation [2], [27]. For instance, SAM-Med2D [2] is trained on 4.6M images and 19.7M masks from public and private datasets, we fine-tune the encoder and decoder of the original SAM to obtain a well-performed SAM-Med2D. SAM-Med3D [27] — a general-purpose segmentation model — is trained on a composite of 94 datasets, achieves generalization across diverse anatomical structures and imaging modalities, and asserts zero-shot transferability to unseen tasks. In contrast, MedSAM-2 [32] builds upon the updated SAM 2 [33] and approaches 3D medical images as video sequences, attaining state-of-the-art performance while preserving strong generalization across a broad spectrum of imaging modalities. VISTA3D [34], meanwhile, is tailored specifically for CT scans, segmenting 127 structures and lesions across highly variable CT data to deliver accurate out-of-the-box results and facilitate straightforward adaptation to unseen anatomical targets.

The other primary approach based on the pretrained foundation model involves automatically generating prompts for the foundation model [2], [3], [26], [27] when inferring new segmentation tasks, enabling the adaptive self-training strategy [6], [7], [28].

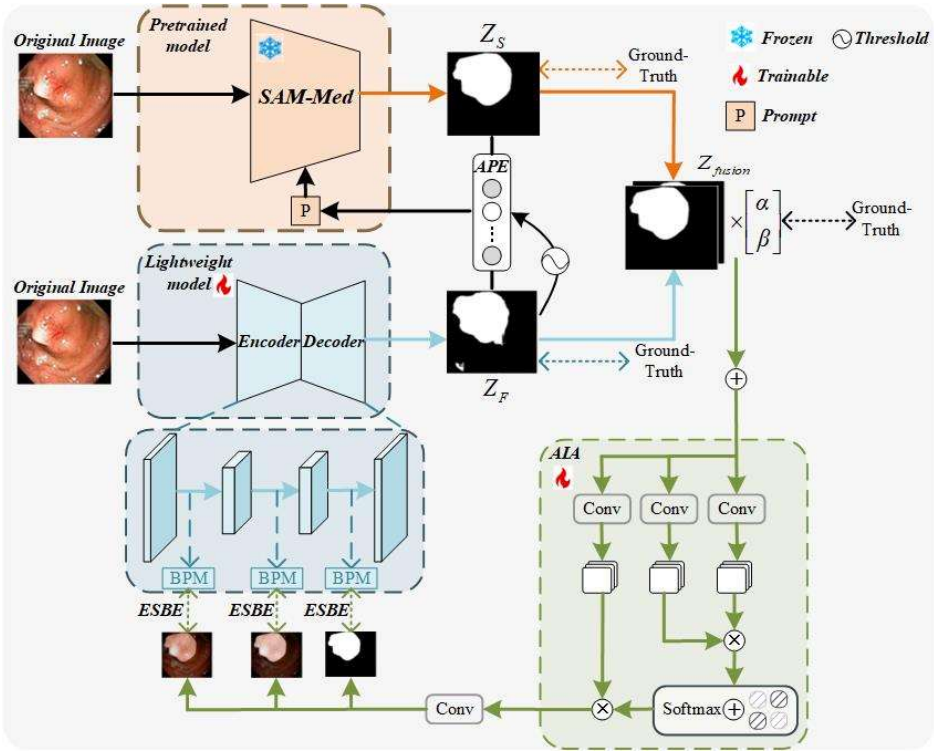


Fig. 1. Overview of the proposed MEI architecture, APE is utilized to dynamically determine high-quality prompt for SAM-Med in each iteration, fully unleashing its potential. ESBE is the ensemble structural boundary enhancement mechanism for precise medical image segmentation and AIA is the integrated self-attention mechanism for initial boundary generation. The pretrained model's branch is denoted as the orange arrow and lightweight model's branch is represented as the blue arrow. The green arrow denotes the process of optimizing a lightweight backbone using the fusion of dual types of results. The double-headed arrow with a dashed line represents the loss propagation.

## III. Methodology

This section introduces the architectural details of MEI, including the backbone, the APE mechanism, the ESBE mechanism, and the objective function. Our method is designed to simultaneously handle both 2D and 3D MedISeg tasks. For clarity of presentation, all modules are described in the 2D format throughout this paper.

### A. Lightweight Backbone

*1) The Architecture of Backbone:* Figure 1 illustrates the overall architecture of our lightweight model branch, which is specifically designed to achieve precise segmentation. The network primarily consists of three key components: encoder, decoder, and bottleneck module. The process begins with an input image $X \in \mathbb{R}^{H\times W\times 3}$, where the encoder's initial feature $E_1$ is first extracted through convolution operations. These features are then fed into the encoder — composed of standard $3\times3$ convolutions followed by batch normalization and ReLU activation function progressively extracting encoder's features $E_i, i \in \{1, ..., l+1\}$. $l+1$ denotes the depth of the model.

The bottleneck module captures multi-scale contextual information $D_{l+1}$ through an Atrous Spatial Pyramid Pooling (ASPP) block [35], and the skip-connection module to bridge the semantic gap between the encoder and decoder:

$$D_i = Conv(Concat(E_i, Dec(D_{i+1}))) \quad i \in \{1, ..., l\} \tag{1}$$

where $Concat$ denotes channel-wise concatenation, $Conv$ represents the $3\times3$ convolution for channel alignment, $Dec$ is the decoder network of the lightweight backbone.

*2) Integration:* Our MEI's final output integrates dual branches consisting of SAM-Med2D generated results $Z_S$, lightweight backbone's outputs $Z_F$. These branches are adaptively fused via a learnable weight matrix $\xi = [\alpha, \beta]^T$ to produce the final segmentation $Z_{fusion}$, with the fusion process defined as:

$$Z_{fusion} = [Z_S, Z_F]\xi \tag{2}$$

### B. Adaptive Prompt for SAM Augmentation

This section describes the APE module from the lightweight model to the pretrained model, as indicated in Figure 1.

*1) SAM-Med2D and Enhancement Goal:* SAM-Med2D is a MedISeg model specialized SAM, adapted by fine-tuning both SAM's image encoder and prompt decoder to accommodate diverse medical imaging data. However, SAM-Med2D's performance heavily depends on prompt quality and it remains a challenging task to obtain accurate and effective prompts. It usually adopts manual generation or iterative refinement methods to optimize SAM-Med2D's prompt, which is inconvenient for the iteration of prompt and whose effectiveness may be limited by human cognition. Therefore, We design the APE mechanism which can automatically iterate and refine

the prompt according to segmentation results of lightweight backbone and SAM-Med.

**Algorithm 1** Adaptive Prompt Enhancement mechanism

**Parameter**:Ground-Truth $Z_G$, Total numbers of epochs $iterations$

set initial prompt = None, S = None, $iter = 1$
for $iter \leq iterations$ do
  get $d = Dice(Z_F, Z_S)$ and $d^{'} = Dice(Z_F, Z_G)$
  if $d^{'} \geq \zeta$:
    Insert $Z_F$ into set $S$
  if $iter \ \% \ k = 0$:
    $Z_F^{'} = Rand(S, 1)$, prompt = $Z_F^{'}$, S=None
  else:
    if $d \geq \eta$: prompt = None
    else: prompt = $Z_F$
  $iter = iter + 1$
**return** prompt

*2) Adaptive Prompt Enhancement Module:* This module can dynamically generate more accurate prompts in real-time based on segmentation results from our lightweight backbone, thereby fully utilizing the potential of prompts and improving SAM-Med2D's performance.

In each iteration, our model evaluates the inconsistency between the results $Z_F$ from the lightweight backbone and the inference result of SAM-Med2D as $Z_S$. The Dice coefficient between the lightweight backbone's result $Z_F$ and $Z_S$ is denoted as $d$.

A hyperparameter threshold $\eta$ is predefined. If $d \geq \eta$, it indicates low inconsistency between the results of the lightweight backbone and SAM-Med2D. Therefore, SAM-Med2D's original prompt are retained with no action, denoted "None". If $d < \eta$, it indicates significant inconsistency and thus $Z_F$ is used as the prompt to enhance SAM-Med2D. Specifically, every $k$ iterations, we first identify all predictions from the lightweight model that achieve a Dice coefficient with the ground-truth above the threshold $\zeta$. From these segmentation results, $Z_F^{'}$ is randomly selected as SAM-Med's prompt. If no result satisfies this condition, the prompt is set to "None". This process can be described as Algorithm 1:

Where $Rand(S, 1)$ denotes the random selection of one element from the set $S$, $Dice(.,.)$ is the calculation process of Dice coefficient. And it should be noted that the initial value of the prompt is set to $None$. And the APE module will downscale $Z_F$ by a factor of 4 through two convolutional embedding layers before the refined prompt is fed into the SAM-Med's prompt encoder.

### C. Ensemble Boundary Enhancement for Lightweight Model

This section details the ESBE module transferring features from the pretrained model to the lightweight model, as shown by the green arrow in Figure 1. Its goal is to optimize the lightweight MedISeg model through the boundary information obtained from SAM-Med2D's inference results. Specifically, the structure boundary $M$ is derived from the fusion of dual results, and $M$ structurally constrains the boundary preservation of the lightweight backbone at the same time.

*1) Adaptive Integration Attention mechanism:* AIA is proposed to effectively fuse the boundary results from dual heterogeneous model branches, including the lightweight backbone and SAM-Med2D. Finally, AIA generates structure boundary $M$ to enhance the lightweight backbone.

The input of AIA is the boundary feature maps corresponding to $Branches = \{Z_S, Z_F\}$ denoted as $\mathbf{F} = \{F_i\}_{i=1}^{|Branches|}$, where $|Branches|$ is the number of branches. $\xi = [\alpha, \beta]^T$ represents their corresponding learnable adaptive weights. Therefore, the weighted feature maps can be represented as $\mathbf{G} = \mathbf{F} * \xi$, where $*$ denotes the Hadamard product.

The cross-attention mechanism [36] is employed to fuse the boundary results of multiple branches. Since the enhancement effect of diverse branches in backbone may vary, we apply three feature maps to respectively serve as $q, k, v$ , resulting in $|Branches|^2$ possible combinations. The attention score matrix for one combination is given by:

$$\mathcal{A} = \|_{l=1}^{|Branches|^2}\Big(\sum_{i=1}^{|Branches|}\sum_{j=1}^{|Branches|}\frac{(W_q^{(l)}\mathbf{G}_i)(W_k^{(l)}\mathbf{G}_j)^T}{\sqrt{d_k}}\Big) \tag{3}$$

where $W_q^{(l)}, W_k^{(l)}, W_v^{(l)}$ are the projection matrices for $q, k, v$, and $\|(.)$ denotes the concatenation along the channel dimension.

For eliminating the effect of self-attention cases where $q, k, v$ all originate from the same branch, a combination mask $diag(-\infty)$ is introduced to the attention score and thus the final attention value $M$ can be computed as follows:

$$M = softmax(\mathcal{A}_{ij} - \infty.\mathbf{1}_{[i=j]})(\|_{l=1}^{|Branches|}(W_v^{(l)}\mathbf{G}_i)) \tag{4}$$

The adaptive matrix $\xi$ is progressively optimized during each iteration, enabling dynamic generation of structure boundary $M$. $\mathbf{1}_{[i=j]}$ takes the value 1 if $i = j$ and 0 otherwise. Finally, we obtain $M_i, i \in \{1, ..., l, l+1\}$ after applying $i-1$ layers of convolutional downsampling to the initial morphological boundary $M_1$.

*2) Structure Boundary Enhancement:* ESBE is designed to enhance the optimization of lightweight backbone by boundary information. The accuracy of boundary preservation directly affects the effectiveness of ESBE. As a result, a new mechanism to enhance the accuracy of boundary preservation by minimizing the semantic discrepancy between the boundary information in the backbone and structure boundaries derived from the AIA mechanism.

A Boundary Preservation Module (BPM) [23] is firstly introduced to extract boundary information from each encoder and decoder layer in the lightweight backbone by employing dilated convolutions with kernels of different sizes. The boundary information is denoted as $B_i$ where $i \in \{1, ..., l, l+1, ..., 2l+2\}$:

$$\{B_i\}_{i=1}^{2l+2} = \{\{\text{BPM}(E_i)\}_{i=1}^{l+1}, \{\text{BPM}(D_i)\}_{i=1}^{l+1}\} \tag{5}$$

TABLE I

COMPARISON BETWEEN OUR MODEL AND EXISTING DEEP LEARNING METHODS ON FOUR PUBLIC DATASETS. THE OPTIMAL RESULTS IN THE TABLE ARE HIGHLIGHTED IN BOLD, WHILE THE SUBOPTIMAL RESULTS ARE INDICATED WITH UNDERLINE. THIS FORMATTING IS CONSISTENTLY APPLIED IN SUBSEQUENT TABLES AND THE IMPROVEMENT OF OUR MODEL COMPARED TO THE SECOND-BEST MODEL IS REPORTED. THE GFLOPS IS GIGA FLOATING-POINT OPERATIONS. THE FPS RESULTS REPORT INFERENCE TIME AND ARE EVALUATED ON GTX 4090 GPU WITH BATCH SIZE OF ONE.

| Method(2D) | Venue | GFLOPs | FPS | Ieee8023 | | Kvasir-SEG | |
|---|---|---|---|---|---|---|---|
| | | | | MIoU(%)↑ | MDSC↑ | MIoU(%)↑ | MDSC↑ |
| UNet | MICCAI'16 | 12.15 | 48.3 | 83.80 | 89.10 | 75.90 | 85.30 |
| BG-Net | MIA'22 | 16.78 | 35.2 | 84.10 | 90.10 | 75.20 | 84.80 |
| BPB-Net | CVPR'20 | 16.44 | 36.1 | 85.50 | 91.10 | 75.70 | 85.20 |
| DSCNet | ICCV'23 | 20.76 | 28.5 | 86.70 | 92.60 | 78.10 | 86.90 |
| FAMNet | AAAI'25 | 25.31 | 18.7 | <u>89.30</u> | <u>93.50</u> | <u>80.30</u> | <u>88.60</u> |
| MADGNet | CVPR'24 | 24.75 | 20.2 | 87.80 | 93.10 | 79.90 | 88.40 |
| MEI | Ours | 30.86 | 12.4 | **90.30** (+1.00%) | **94.50** (+1.00%) | **81.70** (+1.40%) | **90.60** (+2.00%) |
| Method(3D) | Venue | GFLOPs | FPS | CAS2023 | | MM-WHS | |
| | | | | MDSC(%)↑ | MHD↓ | MDSC(%)↑ | MHD↓ |
| 3DUNet | MICCAI'16 | 125.76 | 8.1 | 77.09 | 1.710 | 81.57 | 1.458 |
| nnUNet | MICCAI'18 | 445.34 | 2.3 | 78.25 | 1.687 | 79.11 | 1.639 |
| DSCNet | ICCV'23 | 287.61 | 3.9 | 79.83 | 1.662 | 80.60 | 1.527 |
| LKM-UNet | MICCAI'24 | 279.35 | 4.8 | 77.10 | 1.709 | 77.57 | 1.704 |
| FAMNet | AAAI'25 | 310.74 | 3.2 | <u>80.16</u> | 1.657 | 82.61 | 1.219 |
| MADGNet | CVPR'24 | 308.46 | 3.3 | <u>80.16</u> | <u>1.655</u> | <u>83.16</u> | <u>1.112</u> |
| MEI | Ours | 348.92 | 2.9 | **82.48** (+2.32%) | **1.599** (+3.38%) | **85.73** (+2.57%) | **1.027** (+7.64%) |

$$\text{BPM}(x) = \sigma(Conv(Concat(d_1^1(x), d_1^3(x), d_2^3(x), d_4^3(x), d_6^3(x))) \tag{6}$$

Where $x$ is the input tensor, $d_r^s$ denotes a dilated convolution with a dilation rate of $r$ and a kernel size of $s \times s$, $\sigma(.)$ represents the sigmoid function.

The goal of ESBE can be regarded as minimizing the boundary loss $L_B$ between the boundary features $B_i$ and structure boundaries $M_i$ derived from our model. Therefore, the objective function for ESBE is given by:

$$L_B^i = \begin{cases} L_{CE}(M_i, B_i) & i \in \{1, ..., l, l+1\} \\ L_{CE}(M_{i-(l+1)}, B_i) & i \in \{l+2, ..., 2l+2\} \end{cases} \tag{7}$$

where $L_{CE}(.,.)$ denotes the cross-entropy loss.

### D. Objective Function

The overall optimization objective of our model consists of the boundary loss and two segmentation loss for those branches and a final segmentation loss. We do not emphasize the control of the weights between them because we have already introduced adaptive weight parameters $\xi$ earlier. The final loss is defined as follows:

$$L = \sum_{i=1}^{2l+2} L_B^i + \sum_{Z_i \in \mathbf{Z}} L_{CE}(Z_i, Z_G) \tag{8}$$

where $Z_G$ denotes the ground-truth mask and $\mathbf{Z} = \{Z_S, Z_F, Z_{fusion}\}$.

## IV. EXPERIMENTS

### A. Dataset

We selected MedISeg datasets encompassing multiple organs with varying dimensionalities (2D/3D).

**Ieee8023 (Lung):** This dataset is collected by the University of Montreal. It contains chest CT sequences from 20 COVID-19 patients. 3520 CT slices with labeled lung infection regions were selected and resized to $512 \times 512$. The dataset is publicly available on an open-source platform [37].

**Kvasir-SEG (Rectum):** This is an endoscopic dataset designed for pixel level colon polyp segmentation, comprising 1,000 gastrointestinal polyp images with corresponding segmentation masks in JPG format. All annotations were labeled and verified by senior gastroenterology specialists [38].

**CAS2023 (Vessel):** This dataset comprises MR-modality scans with cerebral artery annotations. The training set includes 100 cases of human brain magnetic resonance angiography (MRA) in NIfTI (.nii.gz) format [39].

**MM-WHS (Artery):** This dataset is a heart segmentation dataset with annotations for a total of 8 heart organs, including the pulmonary artery. It contains 40 MR and CT scans in the training set, and 80 scans in the test set. [40].

### B. Experimental Setup and Evaluation Metrics

**Experimental Setup:** The experiments were conducted with dual NVIDIA GeForce RTX 3090 GPUs, utilizing the PyTorch framework. The dataset was partitioned into training, validation, and test sets at a ratio of 6:2:2. The learnable matrix $\xi$ was initialized as $[1, 1]^T$. Threshold configurations for the

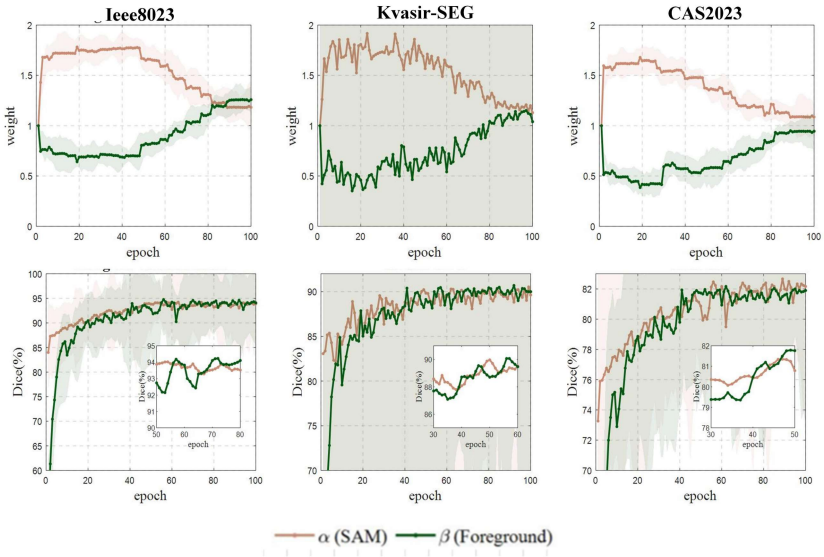


Fig. 2. Weight dynamics and Dice coefficient evolution during 100-epoch training on the Ieee8023, Kvasir-SEG and CAS2023 dataset. The shaded regions in the figure represent variation ranges across three experiments with different random seeds, where pink color represents the SAM-Med branch, green color denotes the lightweight backbone.

TABLE II
QUANTITATIVE COMPARISON OF 2D/3D SEGMENTATION RESULTS ON IEEE8023 AND CAS2023 DATASET AGAINST FINE-TUNED SAM. W/O FT DENOTES TRAINING WITHOUT FINE-TUNING WHILE W FT DENOTES TRAINING WITH FINE-TUNING.

| Method(2D) | Venue | GFLOPs | FPS | Ieee8023 | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | MIoU(%)↑ | MDSC↑ | MPC(%)↑ | MSE(%)↑ | MSP(%)↑ |
| SAM-Med2D(w/o FT) | arXiv'23 | 297.30 | 31.0 | 79.80 | 84.10 | 85.80 | 83.50 | 83.70 |
| SAM-Med2D(w FT) | arXiv'23 | 297.30 | 31.0 | 89.70 | 93.80 | 94.50 | 93.20 | 93.10 |
| nnSAM | arXiv'23 | 481.43 | 12.1 | 87.40 | 92.20 | 93.80 | 91.70 | 91.80 |
| SAMUS | arXiv'23 | 309.74 | 22.2 | 88.40 | 93.70 | 94.80 | 93.10 | 93.00 |
| MedSAM | NAT COMMUN'24 | 312.65 | 21.3 | 88.50 | 93.00 | 94.10 | 92.90 | 92.70 |
| MEI | Ours | 30.86 | 12.4 | **90.30** | **94.50** | **95.30** | **93.90** | **93.50** |
| Method(3D) | Venue | GFLOPs | FPS | CAS2023 | | | | |
| | | | | MDSC(%)↑ | MHD↓ | ACC(%)↑ | AUC(%)↑ | clDice(%)↑ |
| SAM-Med3D(w/o FT) | arXiv'23 | 4821.50 | 3.2 | 70.36 | 1.858 | 86.39 | 78.84 | 75.80 |
| SAM-Med3D(w FT) | arXiv'23 | 4821.50 | 3.2 | 80.09 | 1.635 | 96.41 | 88.42 | 85.38 |
| VISIA3D | arXiv'24 | 5120.82 | 2.9 | 80.25 | 1.618 | 96.52 | 88.57 | 85.56 |
| MedSAM-2 | arXiv'24 | 4675.48 | 3.8 | 81.83 | 1.602 | 96.72 | 88.52 | 86.75 |
| vesselFM | CVPR'24 | 4389.77 | 4.1 | 82.10 | 1.600 | 96.77 | 88.76 | 86.45 |
| MEI | Ours | 348.92 | 2.9 | **82.48** | **1.599** | **96.47** | **88.77** | **86.64** |

APE module vary across different segmentation tasks typically ranging between 0.60 and 0.80.

**Evaluation Metrics:** For 2D MedISeg task, we adopt the MIoU (Mean Intersection over Union), MDSC (mean dice similarity coefficient), MPC (mean per-class accuracy), MSE (mean sensitivity) and MSP (mean specificity). While for 3D MedISeg task, we adopt the MDSC, MHD (mean hausdorff distance), ACC (accuracy), AUC (area under the curve).

### C. Experimental Results and Analysis

*1) Comparison with State-of-the-Art Medical Image Segmentation Models:* As demonstrated in Table I, MEI outperforms baseline models across four 2D/3D MedISeg datasets. Our model achieves 90.30% MIoU and 94.50% MDSC on the Ieee8023 dataset and attains 81.70% MIoU and 90.60% MDSC on the Kvasir-SEG dataset, improving the MIoU and MDSC by average 1.2% and 1.5% compared to the second best model across dual 2D datasets. Moreover, our model reaches 82.48% MDSC and 1.599 MHD on the CAS2023 dataset and scores 85.73% MDSC and 1.027 MHD on the MM-WHS dataset, improving the MDSC and MHD by average 2.4% and 5.5% compared to the runners-up across dual 3D datasets. The results validate our model's generality and specificity in

MedISeg, demonstrating superiority across diverse anatomies and dimensionalities. Our MEI exhibits more generalized segmentation capability compared to conventional deep learning-based lightweight models.

*2) Comparison with Pretrained Foundation Models:* Table II demonstrate our model's capability to fully exploit the potential of the pretrained SAM, achieving obvious improvements beyond conventional fine-tuning approaches. Our model achieves consistent performance gains across both 2D and 3D MedISeg tasks, demonstrating superior 2D metric improvements of +10.5% MIoU and +10.4% MDSC enhancements compared to the pretrained SAM-Med2D baseline without fine-tuning, while in the 3D segmentation tasks it delivers +12.12% MDSC and +13.94% MHD gains versus the pretrained SAM-Med3D. It demonstrates that our MEI can introduce features such as the characteristic of the lightweight model into pretrained models with minimal cost.

*3) Training Dynamics of Adaptive Weights and Performance:* Figure 2 reveals the adaptive weight of the SAM branch (i.e. $\alpha$) exhibit a decreasing trend in training epochs, and those of lightweight backbone demonstrate a progressive increase. This means that the SAM-Med with strong generalization ability takes the dominant position in the early stage of training. As the training progresses, the performance of the backbone improves, and its specificity gradually becomes the dominant factor. Dual branches finally exhibit ascending Dice coefficient, indicating our MEI mechanism enhances SAM-Med through APE and enhances the lightweight model through ESBE sequentially.

*4) Hyperparameter Sensitivity:* To validate the impact of threshold and interval settings in the APE module on model performance, we evaluated cross-dataset performance variations under different thresholds $\eta$, $\zeta$ and the prompt update interval $k$ as shown in Table III.

The experimental results demonstrate that a higher $\eta$ constrains the optimization process of the lightweight model, whereas a lower $\eta$ fails to fully leverage the potential of prompts. A higher $\zeta$ value would render the set $S$ in Algorithm 1 to None, thereby compromising the guaranty of knowledge accuracy in SAM-Med; conversely, a lower $\zeta$ would degrade the quality of the prompt. The moderate $\zeta$ can better optimize the prompt input to SAM-Med, thus enabling its full potential. The prompt update interval $k$ is determined as 10 in all 2D/3D segmentation tasks. The overly small interval $k$ would prevent SAM-Med from learning challenging samples, while an excessively large interval $k$ could lead to deviations during the learning of the challenging samples. Therefore, a moderate interval $k$ can optimally refine the prompts for SAM-Med.

### D. Ablation Study

Table IV demonstrates the respective contributions of APE and ESBE in our framework. The ablation studies across four 2D/3D segmentation datasets reveal that employing either directional enhancement surpasses both the backbone and the fine-tuned SAM-Med2D model in performance.

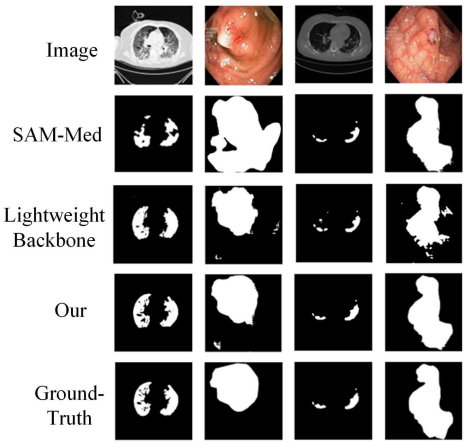


Fig. 3. Columns 1-2 denote where SAM underperforms and lightweight backbone excels, 3-4 show where SAM excels and lightweight backbone underperforms. But our MEI achieves optimal results in the above situations.

As illustrated in Figure 3, our MEI framework achieves superior performance in challenging cases where both the SAM-Med and the lightweight backbone fail to produce accurate results, providing direct evidence for the enhancement effect achieved by our proposed MEI framework.

## V. Conclusion

In this paper, we propose Mutual-Enhanced Integration Between Pretrained SAM and Lightweight Model for MedISeg which can address the underutilization of the pretrained SAM's potential by introducing novel features via the lightweight model. Our framework achieves two key innovations: The APE module to enhance the optimization of SAM-Med by enhancing prompt quality and the ESBE module to enhance the optimization of the lightweight model by providing more accurate boundary information. Extensive experiments on 2D/3D benchmark segmentation datasets demonstrate our model's superior performance across all evaluation metrics.

## References

[1] Goceri, E, Nuclei segmentation using attention aware and adversarial networks. Neurocomputing, 579: 127445, 2024.
[2] Junlong Cheng, Z. D., Jin Ye and Chen, J, SAM-Med2D. arXiv:2308.16184, 2023.
[3] Alexander Kirillov, N. R., Eric Mintun and Mao, H, Segment Anything. arXiv:2304.02643, 2023.
[4] Ronneberger, P. F. and Brox, T, U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597, 2015.
[5] J. Long, E. S. and Darrell, T, Fully Convolutional Networks for Semantic Segmentation. arXiv:1411.4038, 2015.
[6] Yuchen Mao, Y. L., Hongwei Li and Papanastasiou, G, Semi-Supervised Medical Image Segmentation via Knowledge Mining from Large Models. arXiv:2503.06816, 2025.
[7] Li, X. L. Q. W. P. Y. L. Y. Z. Y., N, Segment anything model for semi-supervised medical image segmentation via selecting reliable pseudo-labels. International Conference on Neural Information Processing (ICONIP), 138–149, 2024.
[8] D. Xie, X. T. L. X. L. D., Y. Zhang and Tian, L, BGFE-Net: A Boundary-Guided Feature Enhancement Network for segmentation of targets with fuzzy boundaries. Neurocomputing, 618: 129127, 2025.

TABLE III
SENSITIVITY ANALYSIS ON DIFFERENT THRESHOLDS $\eta, \zeta$ AND PROMPT UPDATE INTERVAL $k$.

| $\eta$ | MDSC(%)↑ | | | | $\zeta$ | MDSC(%)↑ | | | | $k$ | MDSC(%)↑ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Lung | Rectum | Vessel | Artery | | Lung | Rectum | Vessel | Artery | | Lung | Rectum | Vessel | Artery |
| 0.60 | 94.00 | 90.00 | 82.07 | **85.73** | 0.60 | 93.20 | 90.10 | **82.48** | **85.73** | 0 | 93.80 | 90.20 | 82.31 | 85.09 |
| 0.70 | 94.10 | **90.60** | **82.48** | 85.00 | 0.70 | 94.10 | **90.60** | 82.38 | 85.10 | 5 | 94.20 | **90.60** | 82.06 | 84.86 |
| 0.80 | **94.50** | 90.50 | 82.14 | 85.09 | 0.80 | **94.50** | 90.50 | 81.78 | 85.17 | 10 | **94.50** | **90.60** | **82.48** | **85.73** |
| 0.85 | 94.30 | 90.30 | 82.00 | 85.11 | 0.85 | 93.50 | 90.30 | 82.06 | 84.98 | 15 | 93.60 | 90.40 | 81.77 | 84.01 |
| 0.90 | 94.40 | 90.10 | 81.96 | 84.94 | 0.90 | 94.00 | 90.30 | 81.54 | 84.66 | 20 | 94.10 | 90.00 | 81.94 | 84.39 |

TABLE IV
THE IMPACT OF APE AND ESBE ON MODEL PERFORMANCE. WE USE BODY PARTS TO REPRESENT THE FOUR DATASETS HERE.

| Model | Method | | $MDSC(\%)$ | | | |
|---|---|---|---|---|---|---|
| | APE | ESBE | Lung | Rectum | Vessel | Artery |
| Lightweight Backbone | | | 92.30 | 88.10 | 78.16 | 81.79 |
| SAM-Med | | | 93.80 | 89.80 | 80.09 | 84.06 |
| Ours | ✗ | ✓ | 93.10 | 89.50 | 80.07 | 83.78 |
| Ours | ✓ | ✗ | 93.30 | 89.30 | 79.31 | 83.93 |
| Ours | ✓ | ✓ | **94.50** | **90.60** | **82.48** | **85.73** |


[9] G. Allaire, F. J, A level-set method for vibration and multiple loads structural optimization. Computer Methods in Applied Mechanics and Engineering, 2005.

[10] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., Sanchez, C. I, A survey on deep learning in medical image analysis. Medical Image Anal., 42: 60–88, 2017.

[11] R. Srinath, R. S. and Raghavendran, N., Extracting and Classifying 3D Medical Image Texture Characteristics using GLCM and LBP-Based Techniques. 2025 International Conference on Data Science, Agents Artificial Intelligence (ICDSAAI), 2025.

[12] Zhao S X, C. Y.-Z. W. X. X. Z. L. J. Z. D., Li Z D, J, L. Y, SCOAT-Net: a novel network for segmenting COVID-19 lung opacification from CT images. Pattern Recognition, 119: 108109, 2021.

[13] Soltani-Nabipour J, K. A. and B, N, Lung tumor segmentation using improved region growing algorithm. Nuclear Engineering and Technology, 52: 2313–2319, 2020.

[14] E, N. E., S, K. S, Segmentation of lung from CT using various active contour models. Biomedical Signal Processing and Control, 52: 2313–2319, 2020.

[15] Z. Zhou, N. T., M. M. R. Siddiquee, Liang, J, UNet++: A Nested U-Net Architecture for Medical Image Segmentation. arXiv:1807.10165, 2018.

[16] Ozan Oktay, L. L. F., Jo Schlemper, Lee, M, Attention U-Net: Learning Where to Look for the Pancreas. arXiv:1804.03999, 2018.

[17] Hu Cao, J. C., Yueyue Wang, Jiang, D, Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. arXiv:2105.05537, 2021.

[18] Ali Hatamizadeh, Y. T., Nath, V, UN ETR: Transformers for 3D Medical Image Segmentation. arXiv:2103.10504, 2021.

[19] Shaker, H. R. S. K. M.-H. Y., M. Maaz, Khan, F. S, UNETR++: Delving into Efficient and Accurate 3D Medical Image Segmentation. arXiv:2212.04497, 2024.

[20] Hatamizadeh, Y. T. D. Y. H. R., V. Nath, Xu, D, Swin-UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. arXiv:2201.01266, 2022.

[21] Y. Qi, X. Q. Y. Z., Y. He, Yang, G, Dynamic Snake Convolution based on Topological Geometric Constraints for Tubular Structure Segmentation. 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 6047-6056, 2023.

[22] R. Wang, C. J. J. F., S. Chen, Li, Y, Boundary-aware context neural network for medical image segmentation. Medical Image Analysis, 78: 102395, 2022.

[23] H. J. Lee, S. L. H. G. K., J. U. Kim, Ro, Y. M, Structure Boundary Preserving Segmentation for Medical Image With Ambiguous Boundary. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4816–4825, 2020.

[24] Quansong He, J. W. Z. Y., Xiaojun Yao, He, T, A Lightweight U-like Network Utilizing Neural Memory Ordinary Differential Equations for Slimming the Decoder. arXiv:2412.06262, 2024.

[25] Yuntian Bo, L. L., Yazhou Zhu, Zhang, H, FAM-Net: Frequency-aware Matching Network for Cross-domain Few-shot Medical Image Segmentation. arXiv:2412.09319, 2025.

[26] Alec Radford, C. H., Jong Wook Kim, Ramesh, A, Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020, 2021.

[27] Haoyu Wang, J. Y., Sizheng Guo, Deng, Z, SAM-Med3D: Towards General-purpose Segmentation Models for Volumetric Medical Images. arXiv:2310.15161, 2024.

[28] H. Chen, Y. X. Y. Z., Y. Xu, Cui, L, Test-Time Medical Image Segmentation Using CLIP-Guided SAM Adaptation. 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1866–1873, 2024.

[29] Xuran Pan, R. L., Chunjiang Ge, Song, S, On the Integration of Self-Attention and Convolution. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 805–815, 2022.

[30] X. Xu, L.-F. C., Li, Z, 3D Rigid Motion Segmentation with Mixed and Unknown Number of Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43: 1–16, 2021.

[31] M. Lin, J. M. W. W., S. Chen, Wu, J. Instantaneous Energy Consumption Estimation for Electric Buses With a Multi-Model Fusion Method. IEEE Transactions on Intelligent Transportation Systems, 26: 371–381, 2025.

[32] Jiayuan Zhu, Yunli Qi, and Junde Wu. Medical SAM 2: Segment medical images as video via Segment Anything Model 2. arXiv preprint arXiv:2408.00874, 2024.

[33] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, et al. SAM 2: Segment Anything in Images and Videos. arXiv preprint arXiv:2408.00714, 2024.

[34] Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, et al. VISTA3D: Versatile Imaging SegmenTation and Annotation model for 3D Computed Tomography. arXiv preprint arXiv:2406.05285, 2024.

[35] Chen, L., Papandreou, G., Schroff, F., Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. CoRR, abs/1706.05587, 2017.

[36] Chun-Fu Chen, Q. F., Panda, R, CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. arXiv:2103.14899v2, 2021.

[37] Joseph Paul Cohen and Paul Morrison and Lan Dao and Karsten Roth and Tim Q Duong and Marzyeh Ghassemi, COVID-19 Image Data Collection: Prospective Predictions Are the Future, arXiv:2006.11988, 2020.

[38] Jha, D., Smedsrud, P. H., Riegler, M. A, Kvasir-seg: A segmented polyp dataset. MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26, 451–462.

[39] sunhaozhong, Cerebral artery segmentation Challenge (CAS) 2023. https://codalab.lisn.upsaclay.fr/competitions/ 9804, 2023.

[40] Zhuang, X., Li, L., et al., 2019. Evaluation of algorithms for multi-modality whole heart segmentation: An open- access grand challenge. Medical Image Analysis 58.