MELBA Journal Machine Learning for Biomedical Imaging

MCP-MedSAM: A Powerful Lightweight Medical Segment Anything Model Trained with a Single GPU in Just One Day

Donghang Lyu [1*], Ruochen Gao [1*], Marius Staring [1]

1 Division of Image Processing, Department of Radiology, Leiden University Medical Center, Leiden, the Netherlands
***** These authors contributed equally to this work

Abstract

Medical image segmentation involves partitioning medical images into meaningful regions, with a focus on identifying anatomical structures and lesions. It has broad applications in healthcare, and deep learning methods have enabled significant advancements in automating this process. Recently, the introduction of the Segmentation Anything Model (SAM), the first foundation model for segmentation task, has prompted researchers to adapt it for the medical domain to improve performance across various tasks. However, SAM's large model size and high GPU requirements hinder its scalability and development in the medical domain. To address these challenges, research has increasingly focused on lightweight adaptations of SAM to reduce its parameter count, enabling training with limited GPU resources while maintaining competitive segmentation performance. In this work, we propose MCP-MedSAM, a powerful and lightweight medical SAM model designed to be trainable on a single A100 GPU with 40GB of memory within one day while delivering superior segmentation performance. Recognizing the significant internal differences between modalities and the need for direct segmentation target information within bounding boxes, we introduce two kinds of prompts: the modality prompt and the content prompt. After passing through the prompt encoder, their embedding representations can further improve the segmentation performance by incorporating more relevant information without adding significant training overhead. Additionally, we adopt an effective modality-based data sampling strategy to address data imbalance between modalities, ensuring more balanced performance across all modalities. Our method was trained and evaluated using a large-scale challenge dataset, compared to top-ranking methods on the challenge leaderboard, MCP-MedSAM achieved superior performance while requiring only one day of training on a single GPU. The code is publicly available at https://github.com/dong845/MCP-MedSAM.

Keywords

MedSAM, Lightweight, Modality Prompt, Content Prompt, Modality-based Data Sampling Strategy

Article informations

https://doi.org/10.59275/j.melba.2025-4849
Received: 2024-12-10, Published 2025-05-12
Corresponding author: {d.lyu, r.gao}@lumc.nl

1. Introduction

As a key task in the field of medical image analysis, medical image segmentation serves as the foundation for numerous clinical applications, such as disease diagnosis, treatment planning, and surgical interventions (Litjens et al., 2017). Accurate medical image segmentation can precisely identify anatomical structures and pathological regions, leading to more informed medical decisions. Over the last decade, deep learning-based segmentation models like nnU-Net (Isensee et al., 2021) have achieved significant success in medical image segmentation. These models are tailored to specific imaging modalities (e.g., MRI, CT, ultrasound, as shown in Figure 1) and particular diseases (e.g., prostate tumors, pulmonary nodules). This specialization means that different scenarios require different models. Additionally, their optimization for specific datasets limits their generalization across diverse data distributions (Ma et al., 2024a). Therefore, the development of a universal model for medical segmentation becomes increasingly promising in this context, offering the potential to unify approaches across various imaging modalities and clinical applications.

The introduction of the Segment Anything Model (SAM) (Kirillov et al., 2023), the first foundation model for image

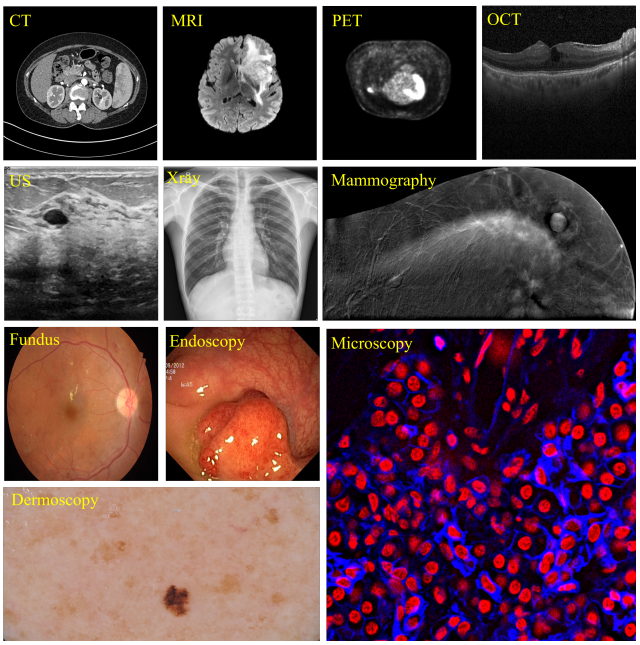

Figure 1: Examples of various medical imaging modalities.

segmentation, marked a significant breakthrough by offering a framework capable of generalizing across a wide variety of natural images. SAM's robust generalization and zero-short capabilities have paved the way for the potential development of general medical segmentation models, enabling adaption across diverse medical image modalities while achieving strong performance on each. Building on SAM's foundation, MedSAM (Ma et al., 2024a) was developed using a large-scale medical dataset covering 10 different imaging modalities and 30 types of diseases. This model has demonstrated superior overall performance compared to traditional, modality-specific specialist models, making a paradigm shift in medical image segmentation. By reducing the need to develop separate models for each imaging modality, it significantly streamlines the entire segmentation process.

Similar to many foundational models, SAM demands significant computational resources, including large-scale GPU clusters and long training time. These requirements limit the applicability and research on the SAM model, especially for smaller research groups and academic institutions that lack enough computational resources. Many previous studies on applying SAM to medical imaging concentrate on freezing the image encoder's weights for direct inference while adding supplementary components for adaptation (Gao et al., 2023; Zhang et al., 2023b), or introducing trainable adapters into the frozen image encoder (Cheng et al., 2023b; Wu et al., 2023). However, the simplicity of these introduced components limits their effectiveness, making it challenging for models to achieve optimal performance, especially when dealing with new imaging modalities or segmentation targets. Recently, some efforts have been made by LiteMedSAM[1], which distills the heavy image encoder into a lightweight version. This significantly reduces model parameters and curtails computational resource consumption, making MedSAM become more accessible for broader research and application. However, this efficiency comes at the expense of segmentation accuracy. Additionally, due to LiteMedSAM's original training strategies, its performance is susceptible to imbalances in imaging modalities within the training data, further limiting its robustness across diverse medical imaging scenarios. Therefore, maintaining optimal overall performance of the MedSAM with a limited computational resource consumption still remains a critical goal.

MedSAM's structure consists of three key components: an image encoder, a prompt encoder, and a mask decoder. While most prior research has primarily focused on optimizing the image encoder and mask decoder, the prompt encoder is usually left unchanged, typically relying on widely used prompt types such as boxes, points, and scribbles. Although these prompts offer valuable spatial hints to help the MedSAM segment targets more accurately, they still require manual effort for annotation, adding to the labor costs. Moreover, determining the optimal quantity and placement of points and scribbles within the given box remains challenging, impacting both model training and final segmentation performance (Cheng et al., 2023a). Therefore, optimizing the prompt encoder with some more effective and stable prompts could be beneficial, as it aims to enhance segmentation accuracy consistently without significantly increasing computational resources and labor costs. In this paper, we introduce **MCP-MedSAM**, which builds on the LiteMedSAM architecture to reduce computational resource costs and accelerate the training process. It incorporates two new types of prompts, a modality prompt and a content prompt, into the prompt encoder, generating sparse embedding and dense embedding representations and integrating them within the mask decoder to further improve segmentation accuracy. For the modality prompt, it is a learnable prompt and its embedding representation aims to take into account the differences among various imaging modalities, enriching the modality-specific information and helping the model adapt to diverse input characteristics. Meanwhile, the embedding representation of content prompt has two types: a sparse one and a dense one, both of which aim to leverage the direct information about the target within the given box but from different perspectives. Additionally, considering the strong zero-shot capability and extensive applications of the CLIP (Radford et al., 2021) model, it is effectively integrated into the processing networks of both prompts to further enhance their representations. We also investigated various data sampling strategies and identified

1. https://github.com/bowang-lab/MedSAM/tree/LiteMedSAM

the most effective one to mitigate the negative impacts brought by the original imbalance. This resulted in enhanced overall segmentation performance, more balanced outcomes across modalities, and a faster training process. In summary, the key contributions of this work are shown as follows:

- **Modality and Content Prompts**: We introduce two new types of prompts into the prompt encoder part of the MedSAM framework: the modality prompt and the content prompt. By generating effective embedding representations and integrating them with the mask decoder, the model's performance across all imaging modalities can be further enhanced without significantly increasing computational resource costs.
- **Lightweight Architecture with Efficient Data Sampling Strategy**: We explore multiple data sampling strategies to identify the most efficient one, aiming to mitigate the effects of data imbalance and further accelerate the training process so that it converges within one day.
- **Efficient Medical Segmentation Model for Broad Accessibility**: After evaluation on a large and diverse set of imaging data, our model demonstrates that it is feasible to achieve high-quality medical image segmentation without the need for extensive computational resources and long training time. This encourages more researchers to adopt and further explore general-purpose medical segmentation models.

2. Related Work

2.1 Medical Image Segmentation

Medical image segmentation has seen substantial advancements through the adoption of deep learning methods. The introduction of the U-Net model by Ronneberger et al. (2015) marks a significant milestone with its U-shaped architecture, effectively combining convolutional layers with symmetric contracting and expanding paths. Its widespread success inspires a host of variants aimed at boosting performance and tackling specific segmentation targets (Isensee et al., 2021; Rahman et al., 2022; Cai and Wang, 2022; Fan et al., 2024; Shu et al., 2024) by incorporating convolutional layers and attention mechanisms into the model design. Among them, nnU-Net (Isensee et al., 2021) stands out as a representative approach, leveraging optimized preprocessing and post-processing to achieve strong medical segmentation performance for each specific medical task. It has been widely adopted in various competitions and real-world applications. For instance, TotalSegmentator (Wasserthal et al., 2023; Akinci D'Antonoli et al., 2025) offers comprehensive and practical solutions for multi-organ segmentation tasks in CT and MRI modalities.

Moreover, transformer modules have been widely integrated into U-shaped architectures for medical segmentation tasks, as transformers excel in extracting global contextual features via their self-attention mechanisms. Notable examples include UNETR (Hatamizadeh et al., 2022) and SwinUNETR (Hatamizadeh et al., 2021), which incorporate transformer modules into the encoder part of the U-Net architecture, yielding enhanced segmentation performance. Likewise, some works (Chen et al., 2021, 2024a; Tang et al., 2024) integrate CNN and transformer modules to leverage the strengths of both architectures and enhance segmentation performance. With the rise of Mamba (Gu and Dao, 2023), recent works (Ruan et al., 2024; Wang et al., 2024b; Liao et al., 2024) have explored integrating the Mamba module into the U-Net architecture for further improvement. Despite these advancements, the aforementioned models are generally tailored to specific segmentation tasks and exhibit limited generalizability across various medical imaging modalities.

2.2 SAM

SAM (Kirillov et al., 2023) is an innovative model that aims to provide a versatile and generalizable solution for segmenting objects in images. There are two key concepts for the success of the SAM model: i) introducing multiple types of prompts, such as bounding boxes, points, and coarse masks, which allows the model to precisely identify and segment the target area; ii) training SAM with a huge amount of data, which enables SAM to adapt to a wide range of segmentation scenarios easily, reflecting its robust zero-shot capability. SAM has gained significant attention, leading to numerous recent works aimed at improving its performance and efficiency. HQ-SAM (Ke et al., 2024) adopts a minimal adaptation approach by introducing a High-Quality output token and fusing global and local features from the image encoder to obtain high-quality features. In this way, HQ-SAM performs better on fine-grained segmentation tasks. SEEM (Zou et al., 2024) introduces more kinds of prompts, including points, boxes, masks, scribbles and text prompts, and learns to deal with them by combining visual and text information in a joint visual-semantic space. This approach enhances segmentation performance and enables zero-shot adaptation to unseen user prompts. Then SAM2 (Ravi et al., 2024) extends the original SAM by enabling both image and video segmentation. By incorporating a memory mechanism, SAM2 can effectively process video data, leading to improved segmentation performance and broader applications. Although SAM models can achieve impressive segmentation performance, it is challenging to train without sufficient computational resources. SAM uses ViT-H (Dosovitskiy et al., 2020) as the image encoder and unifies the input image size to 1024×1024, both of which

contribute to substantial GPU memory usage and make training become time-consuming. Therefore, there are also some works focusing on enhancing SAM's efficiency to make it more suitable for widespread real-world use. MobileSAM (Zhang et al., 2023a) distills the image encoder from ViT-H to a tiny ViT model, significantly reducing the parameter count. Furthermore, EfficientViT-SAM (Zhang et al., 2024) and RepViT-SAM (Wang et al., 2023) replace ViT-H with some lightweight ViT variants, achieving better overall performance with significantly fewer parameters.

2.3 Prior Knowledge in Medical Image Analysis

For a model designed to handle multiple tasks, incorporating prior knowledge, such as features of anatomical structures or modality-specific characteristics, can enhance learning by helping the model recognize internal differences between tasks. DoDNet (Zhang et al., 2021) employs one-hot embeddings to represent different organs, combining these embeddings with image features. This integrated representation is then fed into the output head, enabling DoDNet to segment tumors across various organs. Uniseg (Ye et al., 2023) develops a learnable universal prompt that combines with sample-specific features to create prompts for multiple tasks. Task-specific prompts are then selected based on the task ID, allowing Uniseg to perform segmentation across multiple organs and modalities. Then MedPrompt (Chen et al., 2024b) introduces a self-adaptive prompt block designed to learn and incorporate cross-modal information, enhancing the model's ability to translate effectively across different modalities. Hermes model (Gao, 2024) introduces two kinds of learnable prompts, task-specific and modality-specific, which interact with the model in the bottleneck part to guide the model, enhancing segmentation performance across multiple organs and modalities. Prior knowledge has been effectively used in many models for multi-task processing but remains unexplored in the MedSAM framework for enhancing segmentation across various tasks.

2.4 SAM for Medical Segmentation

Inspired by the success of SAM, some works have begun exploring the utilization of SAM for medical segmentation tasks. MedSAM (Ma et al., 2024a) is proposed to adapt SAM for medical segmentation tasks by training it with a large dataset of medical images. Likewise, SAM-Med2D (Cheng et al., 2023b) fine-tunes the SAM model using a large-scale medical dataset and incorporates a variety of comprehensive prompts, including bounding boxes, points, and masks, rather than relying on just one type of prompt. Med-SA (Wu et al., 2023) extend the SAM structure by introducing adaptors that highly enhance the capabilities for medical applications, enabling it to work with both 2D and 3D medical data. Furthermore, beyond the traditional prompts introduced by SAM, the ScribblePrompt (Wong et al., 2023) model explores the utilization of scribble prompts, resulting in improved overall performance. Similarly, SAT (Zhao et al., 2023) incorporates text prompts into the SAM model, offering contextual medical knowledge about modalities, anatomies, and body regions. With the release of SAM2, researchers have begun exploring its potential in medical segmentation, particularly its application for 3D medical segmentation. MedSAM-2 (Zhu et al., 2024) extends SAM2 for 2D and 3D medical segmentation by training on a large-scale medical dataset and incorporating a self-sorting memory bank to efficiently select informative embeddings, enhancing overall performance.

Furthermore, lightweight MedSAM models can be obtained by applying the techniques from lightweight SAM models and training them on medical data. MedficientSAM (Le et al., 2024) distills the knowledge into an EfficientViT (Cai et al., 2023) image encoder and fine-tunes it with a large-scale medical dataset, while Rep-MedSAM (Wei et al., 2024) chooses to distill knowledge into a RepViT (Wang et al., 2024a) image encoder. DAFT (Pfefferle et al., 2024) also adopts the EfficientViT-SAM structure and introduces a data-aware fine-tuning policy inspired by the mixture of experts (MoE) (Miller and Uyar, 1996) concept to further improve the performance of each modality. Swin-LiteMedSAM (Gao et al., 2024) introduces a tiny Swin Transformer (Liu et al., 2021) as the image encoder and builds skip connections between the image encoder and mask decoder. Furthermore, instead of solely using box prompt, box-based points and scribble prompts are automatically generated based on the provided bounding box. Then, LiteMedSAM-Rep (Yang et al., 2024) uses tiny Swin Transformer as the image encoder and trains the whole model from scratch, subsequently distilling the image encoder into a more lightweight RepViT while keeping the prompt encoder and mask decoder fixed. Although the above methods have made some progress in lightweighting the MedSAM, striking an optimal balance between training efficiency and performance remains a challenge.

3. Materials and Methods

3.1 Datasets

For this study, we used the dataset (Ma et al., 2024b) from the CVPR 2024 competition titled "SEGMENT ANYTHING IN MEDICAL IMAGES ON LAPTOP"[2] to train and test the model. The training dataset comprises over one million paired 2D images and their corresponding segmentation masks across 11 distinct imaging modalities, including Computed Tomography (CT), Magnetic Resonance Imag-

2. https://www.codabench.org/competitions/1847

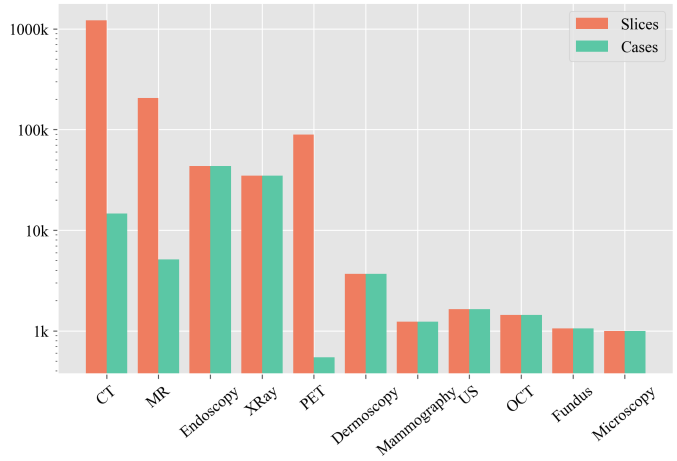

Figure 2: Data distribution across imaging modalities in the training set, with the y-axis displayed on a logarithmic scale for enhanced visualization.

ing (MRI), Positron Emission Tomography (PET), X-ray, Ultrasound, Mammography, Optical Coherence Tomography (OCT), Endoscopy, Fundus imaging, Dermoscopy, and Microscopy. The distribution of the training dataset across these modalities is illustrated in Figure 2. As the testing set was not released after the challenge, we used the competition's validation set as our testing set. Then this testing set comprises 3,278 samples from 9 imaging modalities, excluding Mammography and OCT.

3.2 MCP-MedSAM

Overall, MCP-MedSAM follows the original design of SAM framework, which is composed of three parts: image encoder, prompt encoder and mask decoder, as shown in Figure 3. We introduce two additional prompts into the prompt encoder and design a corresponding network to generate effective representations, enriching the output of prompt encoder. Additionally, the mask decoder is modified to better align with the representations of these prompts. Notably, similar to MedSAM, MCP-MedSAM is designed for 2D medical data and processes 3D medical data slice-by-slice.

While MedSAM-2 delivers strong segmentation performance, its training dataset is limited in both volume and modality diversity. MCP-MedSAM, in contrast, aims for faster convergence and broader modality support, requiring a robust pre-trained image encoder. To achieve this, we build on the lite version of MedSAM, which is trained on a larger, more diverse dataset spanning multiple modalities. A pre-trained tiny ViT[1] image encoder from LiteMedSAM is used to accelerate convergence by providing some prior information. In addition to the traditional box prompt, we incorporate modality and content prompts into the prompt encoder to enhance relevance and robustness. The modality prompt consists of a modality text and an index for extracting the corresponding modality embedding, while the content prompt takes a cropped, resized image from the specified bounding box as input. Furthermore, the embedding representations of these two prompts are integrated into the mask decoder to better fuse modality and content information, resulting in improved segmentation performance. The mask decoder produces three outputs: the target mask, a predicted IoU score, and an additional prediction for the modality class.

3.2.1 Modality Prompt

Considering the unique characteristics of each modality, integrating modality information into LiteMedSAM is advantageous. Consequently, we introduce the modality prompt to enrich the sparse representation output by the prompt encoder. The modality prompt consists of two components: a modality text for generating a text embedding and a modality index i for retrieving a modality-specific learnable embedding from the embedding pool $P \in \mathbb{R}^{N \times F}$. Here, N is the number of modalities, F denotes the embedding length, and modality index $i \in \{1, \ldots, N\}$ corresponds to a specific modality. In this context, we design each learnable embedding to encapsulate modality-specific information after training, while the text embedding provides a feature representation of the modality from a unique perspective. Given prior information of the modality class, we generate the corresponding text input, $\{Modality\}$ *Image*, and pass it through the frozen CLIP text encoder to get text embedding. Although the CLIP model incorporates some medical-related prior information through its pre-trained weights, the training set contains a broader range of modalities and includes many unseen segmentation targets, adding complexity to the task. To address this, an MLP is added after the text encoder to tune and to adjust the channel dimensions simultaneously. On the other hand, a corresponding learnable modality-specific embedding is selected from the modality embedding pool based on the modality index. Then using an MLP to combine these two embeddings allows them to complement each other, creating a more complete and comprehensive modality representation that delivers modality-specific information effectively. The final modality embedding is appended to the sparse embedding. Notably, both MLPs in the modality prompt share the same structure, consisting of two linear layers with a GELU activation function. The input channel size is 512, while the output channel size is 256.

3.2.2 Content Prompt

The prompt encoder typically takes two types of input prompts: a sparse prompt and a dense prompt. The dense prompt serves as an initial coarse mask of the target, gener-

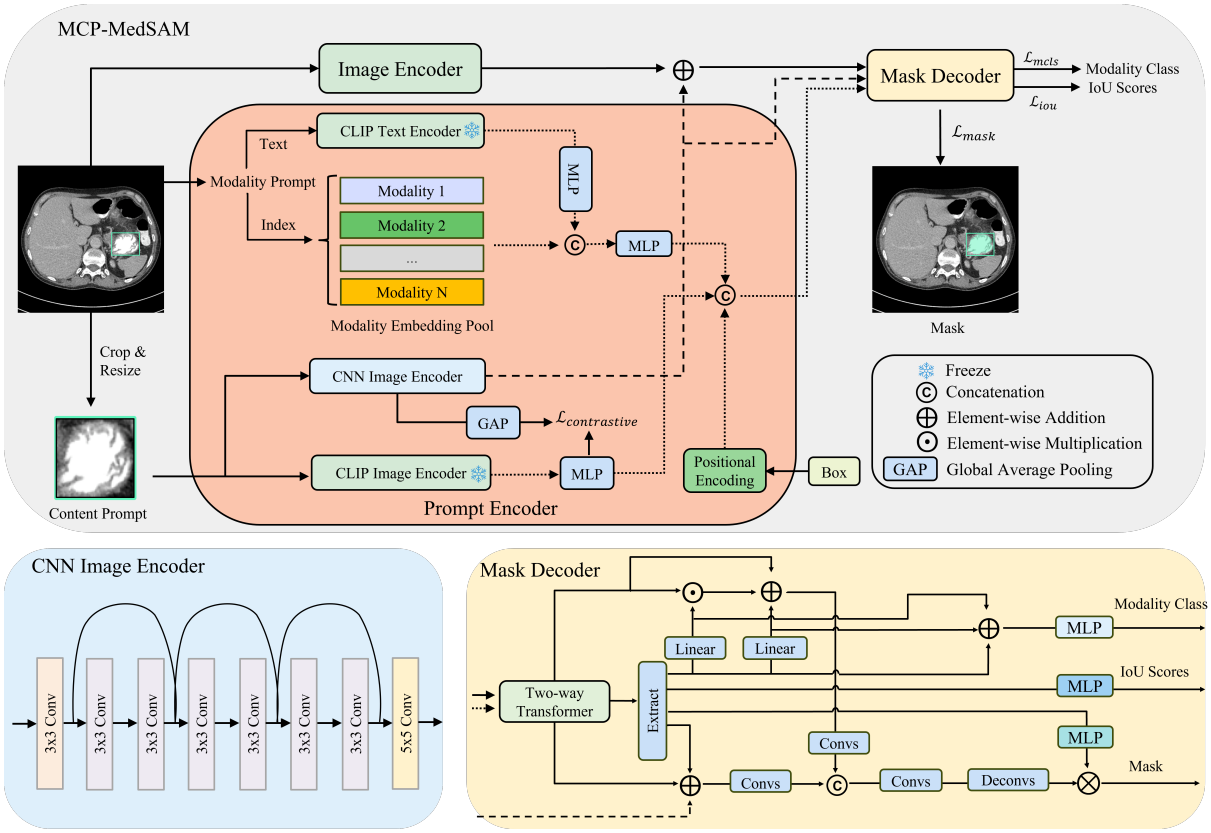

Figure 3: The top section provides an overview of MCP-MedSAM, highlighting our newly introduced content prompt and modality prompt in comparison to the SAM framework. Notably, dotted lines indicate the flow of all sparse embedding representations, while dashed ones represent the direction of dense representation. The detailed architectures of the CNN Image Encoder and the modified Mask Decoder are illustrated in the bottom section.

ating a dense embedding representation through a series of convolutional layers. When the dense prompt is unavailable, the dense embedding representation is instead modeled as a learnable matrix with weights initialized from a uniform distribution. However, the learnable matrix is challenging to train effectively and lacks initial information about the target, while the coarse mask requires significant labor. Therefore, we introduce a content prompt to effectively capture target information within the specified box and generate a dense representation enriched with target features. Additionally, alongside the dense representation, we also extract a sparse embedding containing content information, aiming to enable more direct interaction with other sparse embeddings. Consequently, the final embedding representation of the content prompt has two components: a sparse one and a dense one, each capturing content information from different perspectives. By leveraging both components, the model can achieve a more comprehensive understanding of the content within the specified box, improving its ability to segment the target with greater accuracy. Given a bounding box, the content within the box is cropped and resized into a new shape first. The sparse content representation is derived by processing the reshaped input through a frozen CLIP image encoder, followed by an MLP with the same structure as the MLPs in the modality prompt part (two linear layers with GELU activation function, input channel size is 512 and output channel size is 256). This approach ensures that the sparse content embedding captures the comprehensive content information. For the dense content representation, the image is processed through a CNN image encoder based on the ResNet architecture (He et al., 2016), consisting of multiple convolutional layers with skip connections. With an input channel size of 3, all subsequent convolutional layers maintain a consistent output channel size of 256. Given the small input size, this streamlined CNN network can efficiently extract local detailed features and preserve critical content information about the target. Since both the sparse and dense content embedding representations are generated from the same content, they should exhibit strong similarity. To achieve this, the dense content representation is passed through a global average pooling (GAP) layer to produce a sparse embedding. A contrastive loss is then applied between this sparse representation and the corresponding sparse content embedding to align their learning.

3.2.3 Mask Decoder

In the mask decoder, specific operations are employed to better fuse information from representations of both prompts, enhancing the final performance. After passing the sparse prompts and image embedding through the two-way transformer block, the updated modality embedding is extracted from its corresponding position of the sparse embeddings. In this context, two-way transformer block is a key component of the original SAM (Kirillov et al., 2023) mask decoder. It employs self-attention on tokens and bidirectional cross-attention between tokens and image embeddings, facilitating effective information exchange between image embeddings and tokens, which serve as sparse embeddings. These attention mechanisms also ensure the extracted modality embedding integrates both modality-specific and visual information. Then inspired by the thought of FiLM (Perez et al., 2018), we make the extracted modality embedding pass through two separate linear layers to produce weight and bias embeddings. These embeddings are subsequently combined with the dense matrix output from the two-way transformer through multiplication and addition operations, effectively incorporating modality-specific information. The combined matrix is further processed through two convolutional layers with a skip connection to join in the mask generation process. Furthermore, to better guide the network's learning of the modality prompt within the prompt encoder, we integrate a classification head into the mask decoder, following the same MLP structure as in the prompt encoder but with different input and output channels (two linear layers with GELU activation function, the input channel size is 256 and output channel size is the number of modalities, which is 11 in our experiments). This classification head takes the combination of the extracted embedding along with the corresponding weight and bias embeddings as input and predicts the corresponding modality class.

For the dense content representation, inspired by the global-local feature fusion in HQ-SAM (Ke et al., 2024), we combine it with the dense matrix output from the two-way transformer, which carries global visual information. Furthermore, the sparse content embedding is extracted from the sparse embedding output of the two-way transformer and integrated into this combination as a bias, further enriching the overall information. Finally, similar to the processing of modality information, the resulting matrix is processed through two convolutional layers with a skip connection to refine the features. We concatenate the outputs from both prompt sides and further process them through additional two convolutional layers with a skip connection to fuse the information. The resulting output is then upsampled twice to increase its size and is used to generate the target mask. Notably, the MLPs used for predicting IoU scores and generating segmentation masks all share the same structure, consisting of three linear layers with ReLU activation functions. Both have an input channel size of 256, but their output channel sizes differ: the IoU prediction MLP outputs a single value, while the MLP for mask generation has an output channel size of 32 to align with the feature map.

3.2.4 Loss Function

In this work, except the traditional mask prediction loss $\mathcal{L}_{\text{mask}}$ and IoU (Intersection over Union) score prediction loss $\mathcal{L}_{\text{iou}}$, we also introduce another two loss components: $\mathcal{L}_{\text{mcls}}$ for the modality classification task and $\mathcal{L}_{\text{contrastive}}$ for approaching the similarity between two kinds of content prompts. Therefore, the overall loss function can be represented as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{mask}} + \lambda_2 \mathcal{L}_{\text{iou}} + \lambda_3 \mathcal{L}_{\text{mcls}} + \lambda_4 \mathcal{L}_{\text{contrastive}}. \quad (1)$$

Here, all the λ values are hyperparameters used to tune the importance of each loss component. We set $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = \lambda_4 = 0.01$, emphasizing the primary focus on segmentation while reflecting the auxiliary role of the classification and contrastive tasks in supporting overall performance.

Mask Prediction Loss L_{mask} is the sum of Binary Cross-Entropy (BCE) loss and Dice loss. Given the predicted mask $\hat{M}$ and ground truth mask M, BCE loss evaluates the pixel-wise difference, while Dice loss quantifies the overlap between the two masks. Then their working principles can be shown as follows:

$$\mathcal{L}_{\text{BCE}} = -\sum_{i}^{N} \left(M_i \log(\hat{M}_i) + (1 - M_i) \log(1 - \hat{M}_i) \right), \quad (2)$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_i^N M_i \hat{M}_i}{\sum_i^N M_i^2 + \sum_i^N \hat{M}_i^2}, \quad (3)$$

where N is the total number of pixels and i is the pixel index. Then $\hat{M}_i$ and M_i represent the i-th pixel values of the predicted mask and the ground truth mask, respectively.

IoU Loss $\mathcal{L}_{\text{iou}}$ measures the difference between the predicted IoU score from the model and the ground truth IoU score computed from the overlap between the predicted mask and the ground truth mask, aiming to further improve the accuracy of predicted segmentation mask. We employ a mean squared error (MSE) loss to capture this difference, encouraging precise IoU predictions:

$$\mathcal{L}_{\text{iou}} = \frac{1}{N'} \sum_{i=1}^{N'} \left(s_{\text{iou}}^i - \hat{s}_{\text{iou}}^i \right)^2, \quad (4)$$

where N' is the total number of masks, s_{iou}^i is the ground truth IoU score for the i-th mask, while $\hat{s}_{\text{iou}}^i$ indicates the

corresponding predicted IoU score.

Modality Classification Loss To ensure that modality prompt effectively represents modality-specific information, we introduce an auxiliary modality classification task by employing cross entropy loss function, which is defined as follows:

$$\mathcal{L}_{\text{mcls}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{5}$$

where C denotes the total number of modalities, y_i is the label for each modality class i and $\hat{y}_i$ is the predicted probability for each modality class i.

Contrastive Loss To ensure that pairs of sparse and dense content prompts exhibit the highest similarity, we introduce an auxiliary task that utilizes the contrastive loss from the CLIP (Radford et al., 2021) method. The overall process can be represented as follows:

$$\text{sim}_1 = F_{dc} \cdot F_{sc}^T, \tag{6}$$

$$\text{sim}_2 = F_{sc} \cdot F_{dc}^T, \tag{7}$$

$$\mathcal{L}_{\text{contrastive}} = \left(\mathcal{L}_{\text{ce}}(\text{sim}_1, \mathbf{y}) + \mathcal{L}_{\text{ce}}(\text{sim}_2, \mathbf{y})\right)/2. \tag{8}$$

Here, F_{dc} and F_{sc} are the normalized embeddings of the dense and sparse content prompts, respectively, both having the shape $\mathbb{R}^{B\times C}$, where B is batch size and C denotes the embedding length. sim_1 and sim_2 are two similarity matrices for these two embeddings, each with a shape of $\mathbb{R}^{B\times B}$, where sim_2 is the transpose of sim_1. $\mathcal{L}_{\text{ce}}$ denotes the cross entropy loss function and $\mathbf{y}$ are the collection of labels.

3.3 Training Strategy

In this section, we detail two core components of our training strategy: use of pre-trained models and data sampling strategy. Both are important for accelerating the model's convergence and obtaining a superior overall performance.

3.3.1 Use of pre-trained models

As previously mentioned, many methods freeze pre-trained weights from SAM's image encoder, fine-tuning newly introduced learnable components to improve adaptability of new tasks. Using these pre-trained weights ensures a strong baseline ability for the model. However, instead of following this approach, we opt for a lightweight image encoder, tiny ViT, and use pre-trained weights specifically tailored for the medical image. This allows the image encoder of MCP-MedSAM to fully participate in the training process without being frozen, achieving stronger performance. Additionally, we utilize the pre-trained PubMedCLIP[3] (Eslami et al., 2021) as the frozen CLIP component of MCP-MedSAM, aiming to leverage its strong zero-shot capability and make it provide some medical domain related prior information as well. This variant of CLIP is fine-tuned for the medical domain using the Radiology Objects in Context (ROCO) dataset (Pelka et al., 2018), including multiple modalities from various human regions.

3. https://huggingface.co/kaushalya/medclip

3.3.2 Data sampling strategy

As shown in Figure 2, the distribution of different image modalities in our training dataset is highly imbalanced, primarily due to two factors.

The first factor is varying sizes of public datasets: some modalities, such as CT, MR, and X-ray, have much more publicly available data for AI tasks, resulting in a significantly larger number of training samples. The second factor relates to data dimension. Modalities like CT and MR are in 3D format, enabling the extraction of significantly more slices compared to 2D modalities such as Dermoscopy, Fundus, and Microscopy.

In SAM (Kirillov et al., 2023) and MedSAM (Ma et al., 2024a), data sampling involves iterating over all image slices. In our training set, this approach leads to a significant imbalance: CT slices account for approximately 76% of the dataset, MR slices nearly 13%, while Microscopy slices represent less than 0.1%, making it the least represented modality. Such severe imbalance would negatively impact the model's performance, while training on a large number of slices also results in extended training time.

To address the limitations of existing data sampling methods, we implement a modality-based data sampling strategy, a variant of stratified sampling. This approach prioritizes achieving balance across all modalities, as determined through comparisons with other commonly used data sampling strategies. The details of this strategy are outlined in Algorithm 1. The key point of this approach is randomly selecting a slice from each data case and ensuring that all modalities are evenly sampled within each training batch. This method alleviates the negative impacts of severe data imbalance, allowing all modalities to be trained with an approximately equal number of slices.

4. Experiments and Results

4.1 Experimental Setup

All experiments were performed using Python 3.10 and PyTorch 2.0 on a single NVIDIA A100 GPU with 40GB of memory. The AdamW optimizer was employed with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-3}. A ReduceLROnPlateau scheduler was utilized to decrease the learning rate by a factor of 0.9 every 5 epochs. The batch size was set to 16, and training was conducted

Algorithm 1 Modality-based Data Sampling Strategy

Input: Total number of modalities N, number of cases in each modality $\{C_j\}_{j=1}^{N}$, batch size B
Initialize batch $\mathcal{B} = \emptyset$
for each epoch **do**
  for each batch of size B **do**
    for each sample in the batch **do**
      Select modality $m \sim \mathcal{U}\{0, N-1\}$
      Select case from the chosen modality: $c \sim \mathcal{U}\{0, C_m - 1\}$
      if the selected case is 3D **then**
        Let the shape of the 3D case be $[Z, H, W]$
        Choose slice index $z \sim \mathcal{U}\{0, Z-1\}$
        Extract the slice $\mathcal{S} = \mathcal{D}_{m,c}[z, :, :] \in \mathbb{R}^{H \times W}$
      else
        Extract the 2D case $\mathcal{S} = \mathcal{D}_{m,c} \in \mathbb{R}^{H \times W}$
      end if
      Select mask $k \sim \mathcal{U}\{0, K-1\}$, where K is the number of masks in $\mathcal{S}$
      $\mathcal{B} \leftarrow \mathcal{B} \cup \{(\mathcal{S}, k)\}$
    end for
  end for
end for
Output: Batch $\mathcal{B}$

for a total of 25 epochs. In the final epoch, we set a small learning rate to 5×10^{-5} for an additional fine-tuning of the model. Data augmentation included vertical and horizontal flips, each applied with a 50% probability. Additionally, the Wilcoxon signed-rank test is used to assess the statistical significance of the proposed method in the experiments.

4.2 Evaluation Metrics

Accuracy Metrics Following (Ma et al., 2024a), we adopted the Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD) as evaluation metrics. DSC measures the overlap between two sets, quantifying the similarity between the predicted segmentation and the ground truth. In contrast, NSD evaluates how closely the surfaces of the predicted segmentation align with the ground truth, focusing on surface accuracy. Higher DSC and NSD values correspond to greater segmentation accuracy. To note, DSC and NSD are first averaged within each modality, and then these modality-wise means are averaged to obtain the overall value. Consequently, the standard deviation is also calculated at the modality level.

Efficiency Metrics In order to evaluate the computational efficiency of the model, we established two metrics: GPU training time and CPU inference time. GPU training time is primarily used to assess GPU resource consumption, while CPU inference time focuses on evaluating model performance on edge devices without GPU support, such as laptops or standard CPU workstations. CPU inference time is also part of the competition evaluation criteria[2] for the testing set. The CPU inference time was measured on our local platform with an Intel Xeon(R) W3-2435 @ 3.1 GHz processor and 8GB of memory, where the average time per case was tested in a Docker environment provided by the authors.

4.3 Results

4.3.1 Comparison with Benchmark Models

First, we used the lightweight MedSAM model released by the organizers of the "SEGMENT ANYTHING IN MEDICAL IMAGES ON LAPTOP" competition at CVPR 2024 as our baseline. This model was developed by distilling the image encoder of MedSAM (Ma et al., 2024a) and fine-tuning both the encoder and decoder together using the challenge dataset. Then we compared our model with the top-ranking models from the competition leaderboard: (1) MedificientSAM (Le et al., 2024): MedificientSAM utilizes EfficientViT as its image encoder and increases the input size from 256 × 256 to 512 × 512; (2) DAFT (Ke et al., 2024): DAFT also employs EfficientViT, fine-tuning it on different data subsets based on modality to generate multiple modality-specific models; (3) Rep-MedSAM (Wei et al., 2024): Rep-MedSAM employs RepViT as its image encoder and, after distillation, trains the entire model using the dataset; (4) Swin-LiteMedSAM (Gao et al., 2024): Swin-LiteMedSAM utilizes the tiny Swin Transformer as its image encoder and introduces point and scribble prompts, which are automatically generated based on bounding boxes. (5) LiteMedSAM-Rep: LiteMedSAM-Rep's prompt encoder and mask decoder are initially trained from scratch alongside a tiny Swin Transformer as the image encoder. Then they distill the knowledge from the tiny Swin Transformer to RepViT. Overall, these models mainly focus on modifying the image encoder by distilling knowledge from a large ViT into a lightweight transformer encoder. Table 1 summarizes the segmentation performance of all models, reported in terms of DSC (%) and NSD (%). MCP-MedSAM achieved the best DSC and NSD performance compared to other benchmark models on the testing set. Notably, the segmentation predictions of these benchmark models were reproduced using Docker images provided by the authors, available on Docker Hub[4]. Then Figure 4 presents visualizations of the models' predictions, providing additional details. Overall, our proposed MCP-MedSAM produced seg-

4. https://hub.docker.com/

mentation masks that closely resembled the ground truth, whereas some other models exhibited either a higher degree of over-segmentation or of under-segmentation.

Table 2 demonstrates GPU training time (in hours) and CPU inference time (in seconds) across different methods. MCP-MedSAM achieved the shortest GPU training time (23.8 hours), significantly outperforming all other models. GPU training times were sourced from the respective papers, while the baseline model's training and inference times were not disclosed by the challenge organizers. To note, DAFT has three training stages, but it does not specify the time required for its third stage. Therefore, we estimated the total training time to exceed the combined duration of the first two stages. Additionally, LiteMedSAM-Rep and Rep-MedSAM utilized different types of GPUs (NVIDIA RTX 4090 and NVDIA V100). However, based on Lambda GPU benchmark analysis[5], we find that the A100 40GB is slightly faster than the RTX 4090 (by approximately 1.2×) and significantly faster than the V100 (by about 3.6×). Hence, we estimate that the equivalent training time on an A100 would be approximately 1333 hours for LiteMedSAM-Rep and 52 hours for Rep-MedSAM. As MCP-MedSAM requires only one day (23.8 hours) for training, and LiteMedSAM-Rep and Rep-MedSAM would take an estimated 56 and 2.2 days respectively on an A100, we can still conclude that MCP-MedSAM requires the least training time. For CPU inference time, most methods achieved approximately 1 second. MCP-MedSAM was the slowest among the compared methods, whereas DAFT was the fastest.

4.3.2 Ablation study

In this section, we present a comprehensive ablation study of the key modules in our approach. It is mainly composed of four parts: prompt encoder components of MCP-MedSAM, pre-trained components of MCP-MedSAM, data sampling strategy and training GPU type.

Prompt Encoder Components of MCP-MedSAM Within the prompt encoder, both the modality and content prompt processing networks are composed of two key components. To assess the impact of each, we trained the model under the modality sampling strategy, omitting each component separately. Additionally, we evaluated the model using only one type of introduced prompt, as well as without any introduced prompts, to provide a comprehensive view. To note, when the CNN image encoder branch was excluded, we used a learnable matrix with randomly drawn weights from a uniform distribution as the outputting dense embedding presentation. The detailed results are shown in Table 3. Removing a component from either introduced prompt processing network resulted in a decline in overall performance and might be even worse than removing the entire prompt processing network. Then using only the complete modality or content prompt processing network achieved comparable performance, both of which outperformed the absence of both prompts. Figure 5 further represents the performance of each introduced prompt with some visualizations. Each prompt produced different segmentation results based on its unique perspective, with varying types of errors. Giving both prompts resulted in the best segmentation performance.

Pre-trained Components of MCP-MedSAM As previously mentioned, we incorporated pre-trained weights into the image encoder and CLIP of MCP-MedSAM to improve performance. To support our approach and gain deeper insights into their effects, we conducted several related ablation studies. For the tiny ViT, we tested three scenarios: without pretraining, with pre-trained weights from natural images, and with pre-trained weights from medical images. The CLIP was tested with two options: with pre-trained weights from natural domain, and with pre-trained weights from medical domain. As shown in Table 4, both CLIP and tiny ViT had a better overall performance when using medical domain pre-trained weights compared to those initialized with natural domain pre-trained weights. Furthermore, when Tiny ViT was trained from scratch without any pre-initialized weights, it exhibited the lowest performance among all the options.

Data Sampling Strategy We analyzed three different data sampling strategies: (1) Slice Sampling, used in Ma et al. (2024a), randomly selecting a slice from the whole training dataset; (2) Case Sampling, randomly selecting a slice from each training case; (3) Modality Sampling (see algorithm 1), introducing control over modality balance in each training batch, building on the case sampling approach. The detailed results are shown in Table 5. Among them, slice sampling achieved the highest scores for CT and MR but lagged behind the other two strategies on other modalities. While case sampling and modality sampling showed stronger performance on specific modalities, modality sampling delivered the best overall results with more balanced performance across all modalities.

Training GPU Type To further evaluate the applicability of MCP-MedSAM on smaller GPUs, we conducted an experiment on a mid-range GPU (NVIDIA RTX 6000, 24GB) by reducing the batch size from 16 to 8. The final results achieved a DSC of 86.87 ± 7.53 and an NSD of 88.34 ± 12.07, with a total training time of 54.6 hours. Overall, compared to training on an A100 GPU, segmentation performance decreased and training time increased

5. https://lambda.ai/gpu-benchmarks

Table 1: Accuracy comparison with state-of-the-art methods on the challenge leaderboard, with the best result for each metric highlighted in bold. The † after each metric value indicates a significant difference ($p < .05$) compared to the proposed method.

Models	DSC (%)	NSD (%)
Baseline	$83.81 \pm 15.31^{\dagger}$	$83.26 \pm 22.67^{\dagger}$
LiteMedSAM-Rep (Yang et al., 2024)	$84.51 \pm 10.11^{\dagger}$	$85.03 \pm 17.13^{\dagger}$
Rep-MedSAM (Wei et al., 2024)	$86.19 \pm 7.67^{\dagger}$	$87.97 \pm 11.85^{\dagger}$
Swin-LiteMedSAM (Gao et al., 2024)	$86.78 \pm 8.63^{\dagger}$	$88.44 \pm 12.79^{\dagger}$
MedficientSAM (Le et al., 2024)	$86.20 \pm 8.00^{\dagger}$	$87.65 \pm 11.61^{\dagger}$
DAFT (Ke et al., 2024)	$87.18 \pm 8.29^{\dagger}$	$88.32 \pm 13.41^{\dagger}$
MCP-MedSAM (proposed)	$\mathbf{87.50 \pm 6.91}$	$\mathbf{89.40 \pm 10.37}$

Table 2: Efficiency comparison with state-of-the-art methods on the challenge leaderboard, with the best result for each metric highlighted in bold. Notably, LiteMedSAM-Rep is trained on an NVIDIA RTX 4090, Rep-MedSAM on an NVIDIA V100, and the other models on an NVIDIA A100.

Models	GPU Training Time (hours)	CPU Inference Time (seconds)
Baseline	-	-
LiteMedSAM-Rep (Yang et al., 2024)	1600.0	0.9
Rep-MedSAM (Wei et al., 2024)	188.0	1.3
Swin-LiteMedSAM (Gao et al., 2024)	106.8	2.6
MedficientSAM (Le et al., 2024)	118.5	0.7
DAFT (Ke et al., 2024)	$>$ 42.9	**0.4**
MCP-MedSAM (proposed)	**23.8**	4.6

significantly. This is mainly due to the reduced batch size and lower GPU performance. A smaller batch size leads to more iterations per epoch, resulting in longer training time. More importantly, it can also negatively affect model performance by introducing higher variance in gradient estimation, which makes training less stable and can hinder convergence.

5. Discussion and Conclusion

In accuracy comparisons with the benchmark models (Table 1), MCP-MedSAM achieved the best results and showed a significant statistical difference compared to the other methods. In general, it can be attributed to several key factors: 1) the introduction of two types of prompts offers some valuable cues for target segmentation by integrating information from different perspectives. As shown in Table 3, using either prompt improved overall segmentation accuracy: the modality prompt processing network captures inter-modality differences and unique target characteristics, while the content prompt processing network focuses on features within the bounding box to directly extract information about the segmentation target. The comparable performance achieved with either prompt also indicates the similar importance of modality information and content information contained in the bounding box. Moreover, the components within each prompt processing network complement each other, enriching the overall representation. In contrast, relying on a single network component alone may interfere with the processing of the other prompt, ultimately reducing overall performance; 2) the incorporation of pre-trained components supplies the model with prior knowledge, enhancing its final performance. In Table 4, it is obvious that medical-related initial weights lead to optimal results by incorporating valuable prior information; 3) the modality-based data sampling strategy mitigates the negative effects of data imbalance, leading to a more balanced overall performance compared to the two other strategies. While this slightly lowers the performance of common modalities like CT and MRI, it significantly improves the performance of underrepresented modalities with much fewer training samples, such as PET, as shown in Table 5. Additionally, our proposed method has the lowest standard deviation values, also highlighting the effectiveness of the modality-based data sampling strategy in balancing the performance of multiple modalities.

The visualizations also offered valuable insights for assessing the performance of MCP-MedSAM. In Figure 4, the CT and X-ray samples show multiple overlapping segmentation targets and some of them are small in size, both of which increase segmentation difficulty. In contrast, the other displayed modalities have fewer and larger targets,

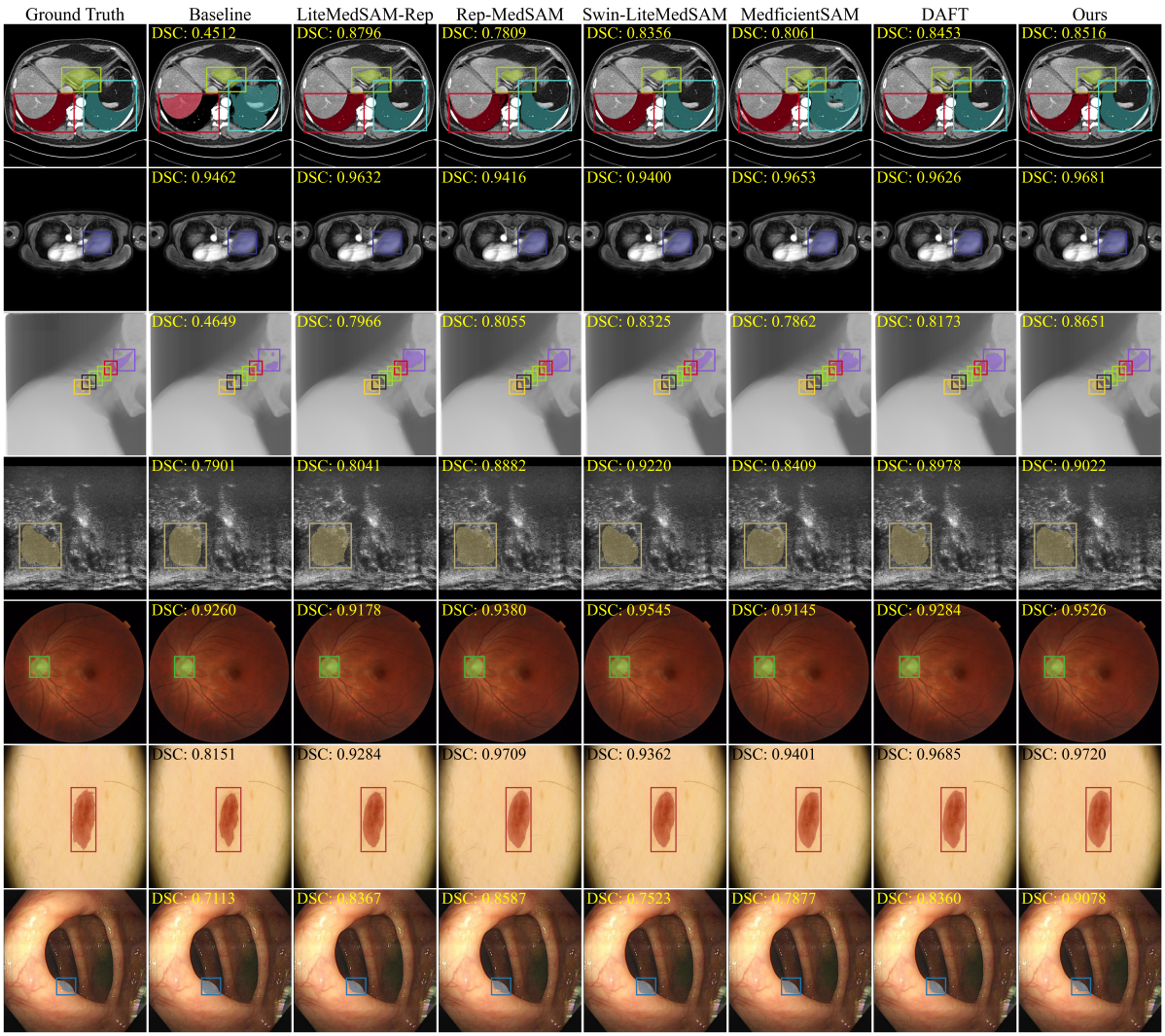

Figure 4: Visualization of multiple modalities yielded by our proposed method and the other benchmark models. The DSC value of each sample is displayed in the upper left corner.

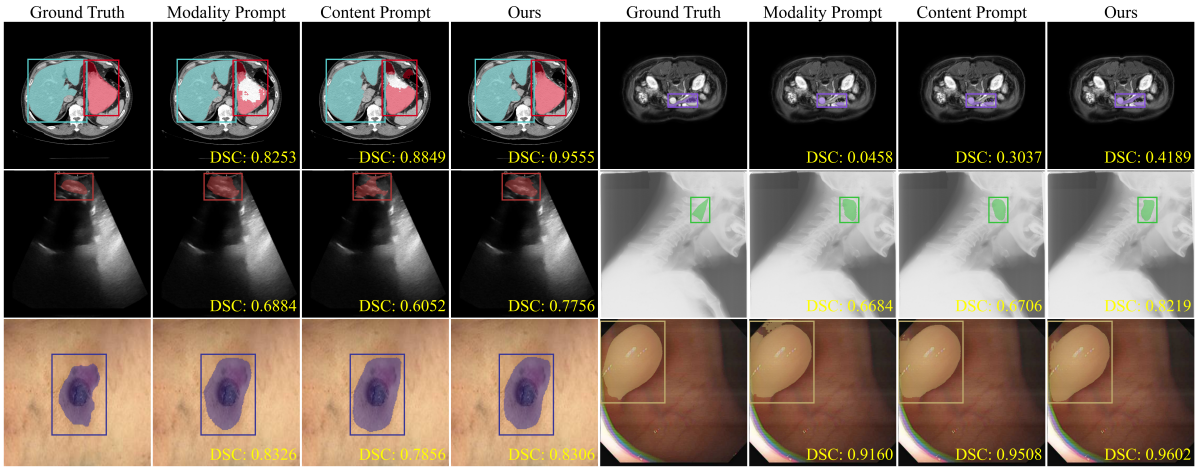

Figure 5: Visualization of multiple imaging modalities generated giving modality prompts alone, content prompts alone, and combined approach.

resulting in lower overall segmentation difficulty. However, noticeable over-segmentation occurred in Endoscopy for many benchmark models, likely due to background features resembling those of the target. Overall, the visualization results align with the quantitative findings in Table 1, and they indicate that small target sizes and complex backgrounds

Table 3: Ablation study of the component of each prompt processing network in the prompt encoder part of MCP-MedSAM model. The checkmark means including the component in the model. And the best result for each evaluation metrics is shown in bold. The † after each metric value indicates a significant difference ($p < .05$) compared to the proposed method.

Modality Prompt		Content Prompt		DSC (%)	NSD (%)
Text CLIP	Modality Embedding	Image CLIP	CNN Encoder		
				$86.36 \pm 8.39^{\dagger}$	$87.64 \pm 13.23^{\dagger}$
	✓	✓	✓	$86.82 \pm 7.97^{\dagger}$	$88.35 \pm 12.72^{\dagger}$
✓		✓	✓	$86.57 \pm 7.21^{\dagger}$	$88.39 \pm 10.91^{\dagger}$
✓	✓			$87.07 \pm 7.19^{\dagger}$	$88.78 \pm 11.07^{\dagger}$
✓	✓		✓	$86.92 \pm 7.83^{\dagger}$	$88.47 \pm 12.31^{\dagger}$
✓	✓	✓		$86.92 \pm 7.59^{\dagger}$	$88.55 \pm 12.02^{\dagger}$
		✓	✓	$87.13 \pm 6.89^{\dagger}$	$88.76 \pm 11.12^{\dagger}$
✓	✓	✓	✓	$\mathbf{87.50 \pm 6.91}$	$\mathbf{89.40 \pm 10.37}$

Table 4: Ablation study of pre-trained components with different weights (natural and medical images) on the MCP-MedSAM model, with the best result for each evaluation metrics highlighted in bold. The † after each metric value indicates a significant difference ($p < .05$) compared to the proposed method.

Method	DSC (%)	NSD (%)
Tiny ViT (No pre-training)	$85.53 \pm 8.67^{\dagger}$	$87.11 \pm 13.35^{\dagger}$
Tiny ViT (Pre-trained on Natural Images)	$86.95 \pm 8.03^{\dagger}$	$88.61 \pm 12.36^{\dagger}$
Tiny ViT (Pre-trained on Medical Images, proposed)	$\mathbf{87.50 \pm 6.91}$	$\mathbf{89.40 \pm 10.37}$
CLIP (Pre-trained on Natural Domain)	$86.99 \pm 7.53^{\dagger}$	$88.44 \pm 12.20^{\dagger}$
CLIP (Pre-trained on Medical Domain, proposed)	$\mathbf{87.50 \pm 6.91}$	$\mathbf{89.40 \pm 10.37}$

could negatively impact segmentation performance, while the proposed MCP-MedSAM represents the best ability to mitigate these effects, demonstrating its robustness and confirming the usefulness of the previously mentioned key factors. Likewise, Figure 5 more clearly demonstrates the benefits of incorporating both introduced prompts into the prompt encoder, as their combination further enriches the feature representation, leading to enhanced performance.

In efficiency comparisons with the benchmark models (Table 2), MCP-MedSAM required the least training time, while the other models took significantly longer time to finish training. For the inference time, the main goal of MCP-MedSAM is to achieve superior segmentation performance without requiring long training time and significant GPU resource consumption, so improving inference speed is not our primary focus. Furthermore, incorporating additional components such as the CLIP component and CNN image encoder will inevitably increase inference time. Several of the benchmark models have adopted different strategies to reduce inference time while maintaining performance, for example, DAFT (Ke et al., 2024) replaced PyTorch with the OpenVINO Runtime[6]. Similarly, Medficientsam (Le et al., 2024) further optimized inference by integrating OpenVINO and leveraging C++ for pre-processing and post-processing optimizations. It could be interesting to adopt such strategies for MCP-MedSAM as well, which will be considered in our future work. Nonetheless, for most clinical scenarios, inference times in the range of seconds are very acceptable, while they can be further reduced by running the models on a GPU.

MCP-MedSAM features a lightweight structure compared to MedSAM (Ma et al., 2024a). While a direct performance comparison is not feasible due to differences in training and testing datasets, the reported performance metrics, such as DSC (ranging from 0.85 to 0.90), indicate comparable performance levels. This highlights MCP-MedSAM's ability to achieve effective medical image segmentation despite its lightweight design. These findings underscore the potential to deliver high-quality segmentation performance without extensive GPU resources, encouraging the development of lightweight and accessible models for advanced medical image analysis. However, MCP-MedSAM adopts the overall design of MedSAM and segments 3D data slice by slice in a 2D manner. This approach is not only time-consuming but also negatively impacts the segmentation performance of 3D samples, as it prevents the model from learning the correlations between slices. With the proposal of the SAM2 framework (Ravi et al., 2024), which achieves superior performance in segmenting both

6. https://docs.openvino.ai/2024/openvino-workflow/running-inference.html

Table 5: Performance comparison across different data sampling strategies, with the performance of each modality detailed. And the best performance for each modality and overall performance is shown in bold. The † after each metric value indicates a significant difference ($p < .05$) compared to the modality sampling strategy.

Modality	Slice Sampling		Case Sampling		Modality Sampling	
	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)
CT	91.00 ± 9.69	93.85 ± 9.70	90.31 ± 8.10	93.55 ± 8.35	90.02 ± 7.98	93.43 ± 8.14
MR	$\mathbf{87.69 \pm 10.70}$	$\mathbf{91.68 \pm 12.41}$	85.73 ± 11.07	89.79 ± 12.06	85.53 ± 11.14	89.81 ± 12.59
PET	66.21 ± 11.38	49.40 ± 30.40	68.06 ± 9.36	52.05 ± 29.13	$\mathbf{73.38 \pm 7.04}$	$\mathbf{61.68 \pm 25.88}$
US	82.50 ± 10.52	87.16 ± 7.20	83.97 ± 11.63	88.71 ± 8.35	$\mathbf{84.77 \pm 9.62}$	$\mathbf{89.61 \pm 6.55}$
X-ray	83.44 ± 7.19	88.24 ± 7.36	$\mathbf{86.33 \pm 5.91}$	$\mathbf{91.00 \pm 6.36}$	85.83 ± 5.93	90.57 ± 6.27
Dermoscopy	93.08 ± 5.21	94.61 ± 4.36	94.58 ± 4.22	96.07 ± 3.23	$\mathbf{94.84 \pm 4.54}$	$\mathbf{96.32 \pm 3.56}$
Endoscopy	93.19 ± 5.12	96.08 ± 3.70	$\mathbf{96.25 \pm 4.11}$	$\mathbf{98.45 \pm 2.19}$	95.17 ± 6.67	97.66 ± 5.08
Fundus	94.57 ± 1.70	96.27 ± 1.52	95.61 ± 1.41	97.21 ± 1.18	$\mathbf{95.77 \pm 1.39}$	$\mathbf{97.35 \pm 1.21}$
Microscopy	76.42 ± 15.67	83.07 ± 12.80	$\mathbf{82.53 \pm 15.79}$	$\mathbf{88.50 \pm 12.74}$	82.17 ± 15.20	88.17 ± 12.28
Average	$85.34 \pm 8.83^{\dagger}$	$86.71 \pm 13.84^{\dagger}$	$87.04 \pm 7.97^{\dagger}$	$88.37 \pm 12.87^{\dagger}$	$\mathbf{87.50 \pm 6.91}$	$\mathbf{89.40 \pm 10.37}$

images and videos, integrating SAM2's working principles with the MCP-MedSAM structure has the potential to enhance segmentation performance for 3D modalities, while also reducing overall inference time. Furthermore, although our results clearly demonstrate quantitative improvement in segmentation performance, important additional work is still needed to explore how these enhancements translate into practical clinical value. For example in radiotherapy (RT) planning, the segmentation of targets such as tumors is critical, as even small errors can significantly impact dose distribution and treatment effectiveness. Therefore, DSC improvements in tumor segmentation are clinically meaningful. However, for certain organs-at-risk (OARs) that are located far from the tumor region—such as the esophagus in some head and neck cancer cases, Mody et al. (2024) showed a low correlation between DSC and dose errors. In such cases, a marginal gain in DSC may not translate into a noticeable clinical difference. Future studies will further investigate the real-world impacts of these segmentation improvements across various clinical scenarios.

In this work, we proposed a lightweight medical segment anything model called MCP-MedSAM, designed to achieve strong overall performance without long training time and large GPU resource consumption. By integrating pre-trained components, the model training process is accelerated, leading to improved performance. Then the introduction of the modality prompt and the content prompts offers valuable diverse information, improving upon the lightweight MedSAM design. Furthermore, a modality-based data sampling strategy ensures that each modality is trained equally, leading to a more balanced overall performance. In conclusion, MCP-MedSAM achieves better overall segmentation performance against top-ranking methods in the challenge[2], demonstrating its effectiveness and potential.

Acknowledgments

This study was supported by the China Scholarship Council (No. 202207720085) and the project ROBUST: Trustworthy AI-based Systems for Sustainable Growth with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO), Philips Research, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023. This study utilized the Dutch national e-infrastructure with the support of the SURF Cooperative using grant No. EINF-6458.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

All data used in the experiments is publicly available.

References

Tugba Akinci D'Antonoli, Lucas K Berger, Ashraya K Indrakanti, Nathan Vishwanathan, Jakob Weiss, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, et al. Totalsegmentator mri: Robust

sequence-independent segmentation of multiple anatomic structures in mri. *Radiology*, 314(2):e241613, 2025.

Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17302–17313, 2023.

Yutong Cai and Yong Wang. Ma-unet: An improved version of unet based on multi-scale and attention mechanism for medical image segmentation. In *Third international conference on electronics and communication; network and computer technology (ECNCT 2021)*, volume 12167, pages 205–211. SPIE, 2022.

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

Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97: 103280, 2024a.

Xuhang Chen, Shenghong Luo, Chi-Man Pun, and Shuqiang Wang. Medprompt: Cross-modal prompting for multi-task medical image translation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 61–75. Springer, 2024b.

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

Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023b.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.

Xiaodong Fan, Jing Zhou, Xiaoli Jiang, Meizhuo Xin, and Limin Hou. Csap-unet: Convolution and self-attention paralleling network for medical image segmentation with edge enhancement. *Computers in Biology and Medicine*, 172:108265, 2024.

Ruochen Gao, Donghang Lyu, and Marius Staring. Swinlitemedsam: A lightweight box-based segment anything model for large-scale medical image datasets. In *Medical Image Segmentation Challenge*, pages 70–82. Springer, 2024.

Yifan Gao, Wei Xia, Dingdu Hu, and Xin Gao. Desam: Decoupling segment anything model for generalizable medical image segmentation. *arXiv preprint arXiv:2306.00499*, 2023.

Yunhe Gao. Training like a medical resident: Context-prior learning toward universal medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11194–11204, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

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

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 574–584, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

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

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

Bao-Hiep Le, Dang-Khoa Nguyen-Vu, Trong-Hieu Nguyen-Mau, Hai-Dang Nguyen, and Minh-Triet Tran. Medficientsam: a robust medical segmentation model with optimized inference pipeline for limited clinical settings. In *Medical Image Segmentation Challenge*, pages 1–14. Springer, 2024.

Weibin Liao, Yinghao Zhu, Xinyuan Wang, Chengwei Pan, Yasha Wang, and Liantao Ma. Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. *arXiv preprint arXiv:2403.05246*, 2024.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

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

Jun Ma, Feifei Li, Sumin Kim, Reza Asakereh, Bao-Hiep Le, Dang-Khoa Nguyen-Vu, Alexander Pfefferle, Muxin Wei, Ruochen Gao, Donghang Lyu, et al. Efficient medsams: Segment anything in medical images on laptop. *arXiv preprint arXiv:2412.16085*, 2024b.

David J Miller and Hasan Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. *Advances in neural information processing systems*, 9, 1996.

Prerak Mody, Merle Huiskes, Nicolas F Chaves-de Plaza, Alice Onderwater, Rense Lamsma, Klaus Hildebrandt, Nienke Hoekstra, Eleftheria Astreinidou, Marius Staring, and Frank Dankers. Large-scale dose evaluation of deep learning organ contours in head-and-neck radiotherapy by leveraging existing plans. *Physics and Imaging in Radiation Oncology*, 30:100572, 2024.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Alexander Pfefferle, Lennart Purucker, and Frank Hutter. Daft: data-aware fine-tuning of foundation models for efficient and effective medical image segmentation. In *Medical Image Segmentation Challenge*, pages 15–38. Springer, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Hameedur Rahman, Tanvir Fatima Naik Bukht, Azhar Imran, Junaid Tariq, Shanshan Tu, and Abdulkareeem Alzahrani. A deep learning approach for liver and tumor segmentation in ct images using resunet. *Bioengineering*, 9(8):368, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491*, 2024.

Xin Shu, Jiashu Wang, Aoping Zhang, Jinlong Shi, and Xiao-Jun Wu. Csca u-net: A channel and space compound attention cnn for medical image segmentation. *Artificial Intelligence in Medicine*, 150:102800, 2024.

Hui Tang, Yuanbin Chen, Tao Wang, Yuanbo Zhou, Longxuan Zhao, Qinquan Gao, Min Du, Tao Tan, Xinlin Zhang, and Tong Tong. Htc-net: A hybrid cnn-transformer framework for medical image segmentation. *Biomedical Signal Processing and Control*, 88:105605, 2024.

Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit-sam: Towards real-time segmenting anything. *arXiv preprint arXiv:2312.05760*, 2023.

Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Repvit: Revisiting mobile cnn from vit perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15909–15920, 2024a.

Ziyang Wang, Jian-Qing Zheng, Yichi Zhang, Ge Cui, and Lei Li. Mamba-unet: Unet-like pure visual mamba for medical image segmentation. *arXiv preprint arXiv:2402.05079*, 2024b.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

Muxin Wei, Shuqing Chen, Silin Wu, and Dabin Xu. Repmedsam: Towards real-time and universal medical image segmentation. In *Medical Image Segmentation Challenge*, pages 57–69. Springer, 2024.

Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: Fast and flexible interactive segmentation for any medical image. *arXiv preprint arXiv:2312.07381*, 2023.

Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

Songxiao Yang, Yizhou Li, Ye Chen, Zhuofeng Wu, and Masatoshi Okutomi. A light-weight universal medical segmentation network for laptops based on knowledge distillation. In *Medical Image Segmentation Challenge*, pages 83–100. Springer, 2024.

Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 508–518. Springer, 2023.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023a.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1195–1204, 2021.

Yichi Zhang, Shiyao Hu, Chen Jiang, Yuan Cheng, and Yuan Qi. Segment anything model with uncertainty rectification for auto-prompting medical image segmentation. *arXiv preprint arXiv:2311.10529*, 2023b.

Zhuoyang Zhang, Han Cai, and Song Han. Efficientvit-sam: Accelerated segment anything model without performance loss. *arXiv preprint arXiv:2402.05008*, 2024.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv preprint arXiv:2312.17183*, 2023.

Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.