Test-Time Adaptation with SaLIP: A Cascade of SAM and CLIP for Zero-shot Medical Image Segmentation

Sidra Aleem[1], Fangyijie Wang[2], Mayug Maniparambil[1], Eric Arazo[3], Julia Dietlmeier[4], Kathleen Curran[2], Noel E. O'Connor[1], Suzanne Little[4]

Abstract

The Segment Anything Model (SAM) and CLIP are remarkable vision foundation models (VFMs). SAM, a prompt-driven segmentation model, excels in segmentation tasks across diverse domains, while CLIP is renowned for its zero-shot recognition capabilities. However, their unified potential has not yet been explored in medical image segmentation. To adapt SAM, to medical imaging, existing methods primarily rely on tuning strategies that require extensive data or prior prompts tailored to the specific task, making it particularly challenging when only a limited number of data samples are available. This work presents an in-depth exploration of integrating SAM and CLIP into a unified framework for medical image segmentation. Specifically, we propose a simple unified framework, SaLIP, for organ segmentation. Initially, SAM is used for part-based segmentation within the image, followed by CLIP to retrieve the mask corresponding to the region of interest (ROI) from the pool of SAM's generated masks. Finally, SAM is prompted by the retrieved ROI to segment a specific organ. Thus, SaLIP is training/fine-tuning free and does not rely on domain expertise or labeled data for prompt engineering. Our method shows substantial enhancements in zero-shot segmentation, showcasing notable improvements in DICE scores across diverse segmentation tasks like brain (63.46%), lung (50.11%), and fetal head (30.82%), when compared to un-prompted SAM. Code and text prompts are available at SaLIP.

1. Introduction

The utilization of Vision-Foundation models (VFMs) has become increasingly prominent in various vision-related tasks, predominantly due to their zero-shot transfer capabilities to various downstream tasks. The Segment Anything Model (SAM) [10] and Contrastive Language-Image Pre-Training (CLIP) [22] have showcased remarkable generalization capabilities in segmentation and recognition, respectively. SAM, in particular, has been trained with a massive dataset of over 1 billion masks, making it highly adaptable to a wide range of downstream tasks through interactive prompts. SAM can be utilized to either segment everything in an image or to segment a specific region based on the prompts. SAM has shown impressive results in a broad range of tasks for natural images but its performance has been subpar when directly applied to medical images [3, 6, 17, 37]. On the other hand, CLIP's training with millions of text-image pairs has given it an unprecedented ability in zero-shot visual recognition.

Both SAM and CLIP have shown remarkable zero-shot transfer capabilities in various downstream tasks for natural images. However, their unified potential in the challenging domain of medical imaging has not yet been explored.

While SAM offers considerable advantages, there are inherent limitations to its application in medical image segmentation. SAM relies on prompts to segment specific regions. This prompt engineering requires domain expertise and manual intervention. However, it is particularly challenging in medical imaging due to the scarcity of high-quality labeled medical data and the need for specialized domain expertise. To address this, several studies have integrated SAM with other foundation models such as GroundingDINO [12] and YOLOV8 [9] to generate bounding box prompts [20] for regions of interest (ROI) [2]. These models are not directly applicable to medical image segmentation. To effectively utilize them for this purpose, they must undergo training with medical datasets containing images paired with their corresponding annotated masks. Their performance is also reliant on the size of training data, which requires careful evaluation and experimentation to achieve optimal segmentation results.

Furthermore, although SAM's capability of automatically segmenting everything in the image is appealing, there are further challenges to its application to medical imaging. One of the main challenges lies in the inherent variability of segmentation tasks. For example, given a liver cancer

[1] ML-Labs, Dublin City University
[2] ML-Labs, University College Dublin
[3] Centre for Applied AI (CeADAR), University College Dublin, Ireland
[4] Insight SFI Centre for Data Analytics, Dublin City University
Corresponding author: sidra.aleem2@mail.dcu.ie

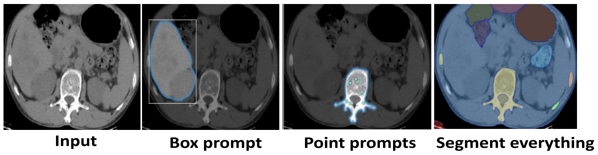


Figure 1. SAM efficiently segments regions based on prompts such as box or point prompts etc. However, such prompt generation needs domain expertise or annotated data, which is not readily available in medical imaging. To overcome this challenge, we use the segment everything mode to get the mask for every part in the image. Then, using CLIP, we select the mask corresponding to the specific organ and use it to generate prompts.

CT image, the segmentation task can vary depending on the specific clinical scenario. One clinician may be focused on segmenting the liver tumor, whereas another may require segmentation of the entire liver along with the surrounding organs. Additionally, clinicians are primarily interested in analyzing specific anatomical organs such as the liver, kidneys, spleen, lesions, etc. It becomes challenging to discern and focus on regions of interest amidst the growing number of segmented areas. Thus, such challenges impede the direct application of SAM to medical image segmentation.

To address the aforementioned challenges, we leverage the combined capabilities of SAM and CLIP and introduce a unified framework called SaLIP, for zero-shot organ segmentation. SAM effectively performs organ segmentation with prompts as shown in Fig. 1, but its effectiveness hinges on domain expertise and annotated data for prompt engineering, which is not readily available in the medical domain. To circumvent these challenges, we adopt the segment everything mode to segment every part in the image and cascade it with CLIP to get the mask for specific organs.

Initially, our framework SaLIP, employs SAM to automatically segment every part within the image. SAM provides exhaustive segmentation, however the resulting masks lack semantic labels. To extract the relevant ROI mask from the pool of generated masks, we first crop the original image according to these masks. This set of cropped regions is passed to CLIP. By employing visually descriptive (VDT) sentences related to the target organ, CLIP then retrieves the corresponding crop in a zero-shot manner [15]. The VDT prompts for CLIP are generated via GPT-3.5 [1]. Finally, the retrieved ROI mask is used for bounding box prompt generation, which is eventually used to prompt SAM to guide the specific organ segmentation. Hence, our framework is training/fine-tuning free and independent of domain expertise or labeled data for prompt engineering. By combining the strengths of SAM and CLIP, our method effectively performs zero-shot medical organ segmentation. We conduct experiments across three diverse medical imaging datasets encompassing MRI scans, ultrasound, and X-ray images to demonstrate the effectiveness of SaLIP.

Our contributions can be summarized as follows:

- We propose a simple unified framework that leverages the combined capabilities of SAM and CLIP for medical image segmentation. We demonstrate that a cascade of these foundation models can improve the zero-shot segmentation accuracy in medical imaging.
- To effectively address the challenges associated with applying SAM directly to medical imaging and to optimize its utilization for medical image segmentation, we propose employing both segment everything and promptable segmentation modes. To the best of our knowledge, we are the first to investigate the utilization of SAM's dual modes for zero-shot medical imaging segmentation.
- Our unified framework SaLIP is adapted fully at test-time for zero-shot medical image segmentation, thereby efficiently alleviating the training costs associated with these foundation models. By leveraging Large language models (LLMs), our method eliminates the need for domain expertise in prompt engineering.

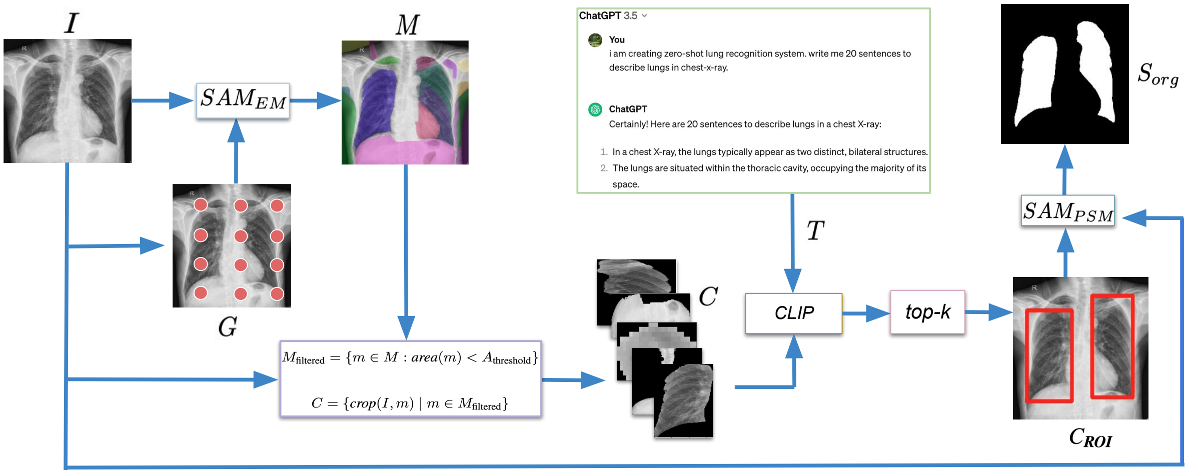
2. Related Work

2.1. Segment Anything Model (SAM)

SAM [10] is a promptable vision foundation segmentation model that aims to segment everything in an image conditioned on different kinds of prompts like bounding boxes and point prompts. It presents a new data engine and portable model for general object segmentation. Given prompts, SAM returns valid segmentation masks. It has three modules: an image encoder, a prompt encoder, and a mask decoder. Masked Autoencoders (MAE) [5], a pre-trained Vision Transformer (ViT) [4] is used as an image encoder. The mask decoder efficiently maps the image embedding, prompt embedding, and an output token to a mask.

2.2. SAM for medical image segmentation

The application of SAM has been investigated within the medical domain. The first line of research focuses on the adaptability of SAM using fine-tuning strategies. MedSAM [14] fine-tunes the SAM mask decoder on large-scale datasets, SAMed [34] adopts a low-rank-based fine-tuning strategy (LoRA) [7], and trains a default prompt for all images in the dataset. Medical SAM Adapter (MSA) [30] uses adapter modules for fine-tuning. These approaches yield promising results, often matching or surpassing state-of-the-art fully-supervised models. Nevertheless, these SAM-based methodologies still require substantial amounts of data for supervised fine-tuning and do not fully leverage the prompting ability. The second line of research focuses on evaluating the performance of few-shot segmentation by prompting SAM to return a specific object segmentation.


I
SAM_{EM}
M
G
$M_{\text{filtered}} = \{m \in M : area(m) < A_{\text{threshold}}\}$
$C = \{crop(I, m) \mid m \in M_{\text{filtered}}\}$
C
T
CLIP
top-k
C_{ROI}
SAM_{PSM}
S_{org}

ChatGPT 3.5

You
i am creating zero-shot lung recognition system. write me 20 sentences to describe lungs in chest-x-ray.

ChatGPT
Certainly! Here are 20 sentences to describe lungs in a chest X-ray:

1. In a chest X-ray, the lungs typically appear as two distinct, bilateral structures.
2. The lungs are situated within the thoracic cavity, occupying the majority of its space.

Figure 2. Illustration of SaLIP: SAM_{EM} segments the input image $\boldsymbol{I}$ using a grid-wise set of keypoints $\boldsymbol{G}$, as prompts to produce part-based segmentation masks $\boldsymbol{M}$. To remove $m \in M$ that corresponds to background, a $A_{threshold}$ is applied on $\boldsymbol{M}$. Subsequently, $\boldsymbol{I}$ is cropped based on $M_{filtered}$ to generate a set of crops $\boldsymbol{C}$, which are then fed into CLIP along with visually descriptive sentences $\boldsymbol{T}$ generated by GPT 3.5. The region of interest crops C_{ROI} are retrieved from CLIP using **argmax**, and **top-k** crops (two in this case) are selected as ROI. The extracted ROI is leveraged to generate bounding box prompts (coordinates shown in red), which are used to prompt SAM_{PSM} to get the specific organ segmentation $\boldsymbol{S_{org}}$ within $\boldsymbol{I}$.

Several recent studies, including [3, 6, 8, 16, 18] have evaluated SAM's capability on different medical image segmentation tasks in the context of zero-shot transfer. However, this prompt generation requires domain expertise or high-quality labeled data.

In contrast, our method is entirely independent of training or domain expertise for prompt engineering. Instead, it effectively adapts SAM to medical imaging segmentation by harnessing the capabilities of both segment-everything and promptable modes, using CLIP as the bridge between the two. Our framework facilitates fully test-time zero-shot organ segmentation in medical imaging.

2.3. CLIP

CLIP [22] is a pre-trained large Vision-Language Model (VLM) known for its strong generalizability and impressive zero-shot domain adaption capabilities. An effective method for adapting CLIP to various domains is through prompt engineering, a process that typically incorporates relevant semantic details related to the specific target task [15]. CLIPSeg [13] extends the CLIP model with a transformer-based decoder that facilitates dense prediction. MedCLIP [29] fine-tunes the CLIP model by separating medical images and texts to expand the available training data exponentially at a low cost. CXR-CLIP [33] improves its performance in chest X-ray classification tasks by fine-tuning the CLIP image and text encoders using samples from image-text and image-label datasets. These methodologies require supervised fine-tuning on medical image-text pairs. Other studies such as [11, 35, 36] have demonstrated that the incorporation of text embeddings learned from CLIP into medical segmentation models achieves state-of-the-art results. However, these medical image-text pairs are collected under guidelines and with the support of domain experts.

3. Methodology

In this section, we first review SAM and CLIP in Sec. 3.1.1 and Sec. 3.1.2. Subsequently, we explain our unified framework SaLIP in Sec. 3.1.3. Our framework is illustrated in Fig. 2.

3.1. Preliminaries

3.1.1 SAM

SAM is a prompt-driven segmentation foundation model. It consists of three main components: an image encoder, a prompt encoder, and a lightweight mask decoder. We denote an input image as $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$ and an input visual prompt as $\boldsymbol{P} \in \mathbb{R}^{N}$, where $\boldsymbol{H} \times \boldsymbol{W}$ are the spatial dimensions and $\boldsymbol{N}$ is the number of prompts. The image encoder is a MAE [5] pre-trained Vision Transformer (ViT) [4]. It encodes an image into dense features $F_{SAM} \in \mathbb{R}^{\frac{H}{16} \times \frac{Q}{16}}$. The prompt encoder encodes prompts $\boldsymbol{P}$ into sparse prompts Q_{sp}. $\boldsymbol{P}$ can either be sparse, such as points, boxes, or text, or dense, like masks. The points and boxes are represented by positional encodings [27] summed with learned embeddings for each prompt type. Currently, SAM does not directly process text prompts and the text-to-mask task is still in its exploratory stages and is not entirely robust [10].

The mask decoder efficiently maps the image features F_{SAM}, Q_{sp} , and an output token to a mask. It uses prompt self-attention and cross-attention in two directions (prompt-to-image embedding and vice-versa) to update all embeddings. After running two blocks, a multilayer perceptron (MLP) maps the output token to a dynamic linear classifier, which then computes the mask foreground probability at each image location.

SAM can operate in two distinct modes: segment everything mode (*SAM**EM***) and promptable segmentation mode *SAM**PSM***. The former can segment everything in the image without relying on externally provided prompts. Instead, a grid of keypoints is generated on the image, and these grid-wise set of keypoints are used as prompts to segment everything in the image. The latter segment a specific set of ROIs based on the prompts given to SAM.

Our framework utilizes both modes of SAM with CLIP as a bridge between them.

3.1.2 CLIP

Using contrastive pre-training on large image-text datasets, CLIP performs image classification. CLIP aligns image and text modalities within a shared embedding space. After pretraining, CLIP directly performs image classification on the target dataset without any fine-tuning. For an image $I \in \mathbb{R}^{H \times W \times C}$, where $\boldsymbol{H} \times \boldsymbol{W} \times \boldsymbol{C}$ denotes spatial dimension, the vision encoder $\boldsymbol{f}$ maps $\boldsymbol{I}$ into a joint embedding space to get the image features $E \in D$ with dimension $\boldsymbol{D}$. During inference, a prompt template such as *'A photo of classname'* is used to generate sentences for K different classes and passed through the text-encoder to yield classifier weight matrix $W \in \mathbb{R}^{D \times K}$. Prediction probabilities are then calculated by multiplying image feature $\boldsymbol{f}$ and $\boldsymbol{W}$ and applying a softmax function.

In this work, to construct textual prompts for CLIP, we use ensembles of visually descriptive (VDT) information for each class [15].

3.1.3 SaLIP

In this section, we provide a detailed overview of our unified framework for zero-shot organ segmentation. First, we explain how *SAM**EM*** generates masks for every part within the image, followed by an explanation of how CLIP retrieves the relevant ROIs from the pool of generated part-based masks. Finally, we illustrate how we leverage retrieved ROIs to create prompts for *SAM**PSM***.

To generate part-based segmentation masks from the image, we use *SAM**EM***. It generates an extremely exhaustive prediction of nearly any object or part in the images. It takes a grid-wise set of keypoints $\boldsymbol{G} \in \mathbb{R}^{g^2 \times 2}$ as input, where $\boldsymbol{g}$ is the point number along one side of the image. Then the masks are generated by *SAM**EM*** by prompting it with a set of grid-wise key points.

$$\mathbf{M} = \mathbf{SAM_{EM}}(\mathbf{I}, \mathbf{G}) \tag{1}$$

where $I \in \mathbb{R}^{3 \times H \times W}$ is the input image, and $M \in \mathbb{R}^{N \times H \times W}$ is the set of all the part-based generated masks. $\boldsymbol{N}$ refers to the number of masks and $\boldsymbol{H} \times \boldsymbol{W}$ is the spatial dimension.

The process of generating part-based masks with *SAM**EM*** is greatly influenced by the selection of hyperparameters utilized for the SAM mask generator module. To streamline and achieve optimal part-based segmentation, we use a random search for optimal hyper-parameters for *SAM**EM*** using five randomly selected images. The combination of hyper-parameters that yields the highest DICE score is used as the final configuration of *SAM**EM*** to generate part-based masks for the entire dataset.

Following this, the next step in the pipeline is to extract the ROI mask from $\boldsymbol{M}$ using CLIP. To accomplish this, we first utilize $\boldsymbol{M}$ to crop $\boldsymbol{I}$, thereby producing a series of crops, each corresponding to a mask in $\boldsymbol{M}$. As *SAM**EM*** generates masks for every element within the image, there arises the possibility of predicting a mask corresponding to the background. In such instances, the resulting crop has spatial dimensions identical to $\boldsymbol{I}$. Consequently, when these crops are subsequently passed to CLIP, there is a risk of miss-classification, as CLIP may perceive them as ROI due to the presence of relevant region in $\boldsymbol{I}$ as discussed in Sec. 4.4.3. To mitigate this issue, instead of directly forwarding the entire set of masks $\boldsymbol{M}$ to CLIP, we first filter out the masks $m \in M$ that potentially correspond to the background region using area-based filtering on each mask within $\boldsymbol{M}$. To determine the optimal threshold for area-based filtering, we perform a random hyperparameter search within the space defined by the areas of masks in $\boldsymbol{M}$. This search is carried out simultaneously with the hyper-parameter optimization process for *SAM**EM***, using the same methodology discussed above. The area filtering is conducted, and set crops are generated as follows:

$$\mathbf{M}_{\text{filtered}} = \{\mathbf{m} \in \mathbf{M} : \textit{area}(\mathbf{m}) < \mathbf{A}_{\text{threshold}}\} \tag{2}$$

$$\mathbf{C} = \{\textit{crop}(\mathbf{I}, \mathbf{m}) \mid \mathbf{m} \in \mathbf{M}_{\text{filtered}}\}, \tag{3}$$

where $A_{threshold}$ is the value of area achieved via hyper-parameter search used for filtering $\boldsymbol{M}$ based on area, $\boldsymbol{m}$ is a mask from $\boldsymbol{M}$, $M_{filtered}$ is the set of masks after removing the $\boldsymbol{m}$ corresponding to background. $\textit{crop}(\mathbf{I}, \mathbf{m})$ denotes the operation of cropping $\boldsymbol{I}$ according to $m \in M_{filtered}$ The set of generated crops $\boldsymbol{C}$ along with the textual prompts are passed to CLIP to select the crop corresponding to ROI. To construct textual prompts for CLIP, we use prompt ensembling, a technique that constructs several sentences for each class and subsequently averages the classification vectors. We use prompt ensembles of visually descriptive (VDT) information for each class [15]. The VDT sentences are generated via GPT 3.5 and passed through CLIP to get the text embeddings and averaged to obtain a single text prototype $\boldsymbol{W_T}$ for the organ under consideration. Now all the image crops in $\mathbf{C}$ are passed through CLIP's vision encoder to obtain vision embeddings $\boldsymbol{E}_c$. Subsequently, the mask corresponding to ROI is computed as:

$$\mathbf{C_{ROI}} = \text{topk}\left(\arg\max_{\mathbf{c}\in\mathbf{C}} \mathbf{S}(\boldsymbol{E}_\mathbf{c}, \boldsymbol{W}_T)\right) \tag{4}$$

where $S(E_c, W_T)$ represents a similarity function which computes cosine similarity between any embeddings $\boldsymbol{E}_c$ of any crop $\boldsymbol{c} \in \boldsymbol{C}$ and the text embeddings $\boldsymbol{W_T}$. $\boldsymbol{k}$ denotes the number of ROIs and varies depending on the number of ROIs in the image. $\boldsymbol{C_{ROI}}$ is the mask corresponding to ROI.

Finally, we compute the bounding box prompts using the minimum and maximum $\boldsymbol{X}$, $\boldsymbol{Y}$ co-ordinates of the retrieved $\boldsymbol{C_{ROI}}$ and use it to prompt $\boldsymbol{SAM_{PSM}}$ as:

$$\mathbf{S_{org}} = \mathbf{SAM_{PSM}}(I, P) \tag{5}$$

where $P \in \mathbb{R}^{k\times 4}$ is the bounding box computed from $\boldsymbol{C_{ROI}}$, $\boldsymbol{N}$ is the number of box prompts which varies according to ROI and $\boldsymbol{S_{org}}$ is the final segmentation for ROI.

4. Experiments

4.1. Datasets and Metrics

We assessed our method across three diverse medical imaging modalities, encompassing two datasets focusing on single-organ segmentation and one more challenging dataset requiring the segmentation of two distinct organs. Calgary-Campinas (CC359) [25] is a multi-vendor (GE, Philips, Siemens), multifield strength (1.5, 3) magnetic resonance (MR) T1-weighted volumetric brain imaging dataset. It has six different domains and contains 359 3D brain MR image volumes, primarily focused on the task of skull stripping. The HC18 [28] consists of 2D fetal head ultrasound images obtained throughout all trimesters of pregnancy. These images have been annotated with biometrics by experienced medical experts. From this dataset, a subset of 200 images is selected for testing purposes. X-ray Masks and Labels [19] consists of 800 2D chest X-ray images, each accompanied by its corresponding mask for lung segmentation. We use the DICE score (DSC) and mean intersection over union (mIoU) as our evaluation metrics.

4.2. Implementation Details

We employed ViT-H, a variant of SAM, and ViT-L/14 trained in CLIP by OpenAI. For CLIP, the visually descriptive textual sentences are generated using GPT 3.5, the details can be found in the supplementary material, Sec. 9. We implemented our framework in PyTorch [21] with SAM codebase [1]. All experiments are performed on a desktop computer with the Ubuntu operating system 20.04.6 LTS with CUDA 11.6, NVIDIA GeForce RTX 3090 GPU. For reproducibility, a random seed is set to 1234.

4.3. Results and Analysis

We evaluated our proposed method against U-Net [23], ground truth-aided SAM (GT-SAM), and un-prompted SAM. U-Net is widely used in medical image segmentation, and it is fine-tuned for all three of our datasets separately. GT-SAM is an upper bound in which box prompts for SAM are directly derived from the ground truth. As our method does not utilize ground truth for prompt generation, to ensure a fair comparison and simulate the real-world medical imaging scenarios without annotated data, we employ an un-prompted version of SAM. In this version, SAM is not provided with prompts from ground truth, rather it utilizes its default prompt embedding.

The quantitative results are shown in Tab. 1. The difference in performance among all methods can be attributed to the prompts utilized for SAM, highlighting the significant dependency of SAM's performance on the prompts employed. For CC359 [25], our method achieves an average of 0.94 DSC, significantly outperforming the un-prompted SAM's average DSC of 0.31. When evaluated for lung segmentation, our approach elevates the performance from an initial DSC of 0.31 with unprompted SAM to 0.83. Our method achieves an average DSC of 0.81 for segmenting the fetal head on HC18 [28], achieving a 26% increase compared to the unprompted SAM's average DSC of 0.55. The low performance of the un-promoted SAM is because it does not utilize any prompts, which leads to its inability to segment the ROIs. This demonstrates that SAM's ability to effectively segment is strongly reliant on the prompts. Furthermore, the qualitative analysis presented in Fig. 3, illustrates that the un-prompted SAM fails to accurately perform organ segmentation, as it generates a general segmentation mask for the regions in the image rather than delineating specific organs. Hence, SAM's applicability to medical imaging scenarios is limited, where obtaining domain expertise and annotated data for prompt engineering is challenging. On the other hand, GT-SAM achieves high DSC of 0.95, 0.94, and 0.91 for brain, lungs, and fetal head segmentation respectively. This high performance is attributed to GT-SAM's use of perfect prompts extracted directly from the ground truth. In contrast, our method performs fully test-time zero-shot organ segmentation without relying on external prompts or domain expertise, demonstrating its ef-

[1] https://github.com/facebookresearch/segment-anything

ROI	Dataset	U-Net		GT-SAM		Un-prompted SAM		Ours	
		DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU
Brain	GE 1.5	0.98	0.93	0.95	0.91	0.33	0.29	**0.92**	**0.87**
	Philips 1.5	0.97	0.95	0.96	0.93	0.41	0.31	**0.94**	**0.85**
	Philips 3	0.95	0.92	0.93	0.89	0.40	0.39	**0.89**	0.80
	Siemens 1.5	0.97	0.95	0.95	0.91	0.39	0.26	**0.90**	**0.81**
	Siemens 3	0.98	0.92	0.96	0.90	0.41	0.32	**0.93**	**0.85**
Lungs	X-ray	0.98	0.95	0.94	0.90	0.47	0.31	**0.83**	**0.76**
Fetal head	Ultrasound	0.95	0.91	0.95	0.91	0.55	0.40	**0.81**	**0.72**

Table 1. Comparison of our method with other baselines. Our method significantly outperforms un-prompted SAM, without using domain expertise or annotated data for prompt engineering. **Note:** GT-SAM uses the prompts extracted from ground truth.

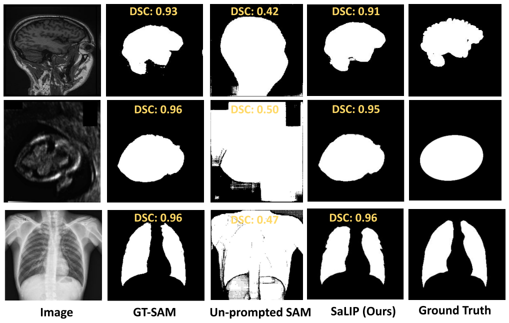


Figure 3. GT-SAM is the upper bound, un-prompted SAM, SaLIP (ours). The text in yellow refers to the DSC with each method respectively.

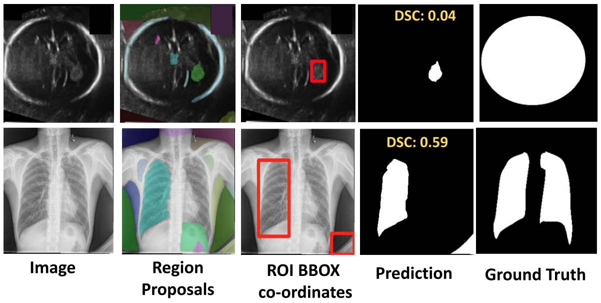


Figure 4. SAM failure cases: First row: SAM_{EM} fails to generate a mask for the fetal head, resulting in miss-classification by CLIP. Second row: SAM_{EM} generates a mask for the right lung but fails to generate one for the left lung, leading to CLIP retrieving the wrong crop.

fectiveness and versatility. To evaluate our method in comparison to GT-SAM, it is important to highlight that GT-SAM benefits from perfect prompts extracted directly from ground truth and is the upper bound. In contrast, our method operates in a zero-shot manner completely independent of ground truth or any domain expertise for prompt engineering. Despite this, our method still achieves results comparable to GT-SAM. This demonstrates the effectiveness and adaptability of our approach in the context of organ segmentation in medical images, where access to domain expertise and perfect annotated data is either limited or impractical. To further demonstrate the effectiveness of our approach, the qualitative results are shown in Fig. 3.

4.3.1 Failure Cases and Future Work

Although our method demonstrates effective performance, through an in-depth analysis and exploration, we have identified two sets of limitations: one at the SAM level and the other one at the CLIP level.

SAM part-based segmentation: This refers to the instances where SAM_{EM} fails to generate a mask for the ROI as shown in Fig. 4. Among our three datasets, this issue is particularly prominent in ultrasound and X-ray images. Due to the nature of how ultrasound images are captured, inherent limitations in fetal ultrasound images are very common, such as acoustic shadows, speckle noise, and obscured boundaries. These characteristics pose challenges to the accurate generation of masks for the fetal head by the SAM_{EM}. These issues often arise from sub-optimal selection of hyperparameters for SAM's mask generation process. To tackle this, we have implemented an automated hyperparameter search for SAM_{EM} hyperparameters. This automation significantly mitigated the problem.

CLIP mask retrieval: There are instances where SAM_{EM} generates masks corresponding to ROIs, but CLIP fails to retrieve them. Such issues arise due to the generation of multiple masks for a single region by SAM_{EM} and impact the datasets with multiple ROIs, lungs in our case. CLIP in some cases retrieves the masks corresponding to the same lung region as shown in Fig. 5.

For extracting lung crops from the pool of masks generated by SAM_{EM} via CLIP, we use a single set of visually descriptive prompts that characterize both lungs in a chest

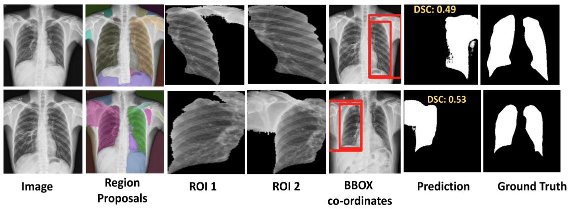


Figure 5. CLIP failure cases: SAM_{EM} generates multiple masks for both ROIs (left and right lung). First row: CLIP while correctly recognizing the left lung, identifies a second mask for the same lung region and fails to retrieve the one for the right lung. Second row: CLIP does not retrieve the left lung crop.

X-ray (supplementary Sec. 9.1). To address the challenge of the same region mask retrieval by CLIP, we experimented with a separate set of prompts for the left and right lungs and evaluated the impact. For more details, please refer to the supplementary material (Sec. 9.2). However, our results demonstrated that CLIP shows limited performance in precise localization and recognition tasks and lacks semantic knowledge in distinguishing objects based on their spatial alignment (i.e., left and right). Consequently, employing separate sets of prompts to describe organs based on their spatial alignment does not mitigate the issue. The results are presented in Tab. 2. In contrast to using separate prompts, our approach of utilizing a single set of prompts describing both lungs achieves 0.83 DSC, thereby outperforming the separate prompts, which achieve 0.67 and 0.28 DSC for the left and right lung, respectively.

	Right Lung	Left Lung	Both (Ours)
DSC	0.67	0.28	**0.83**

Table 2. CLIP performance with separate prompts for left and right lungs, and combined prompts.

Recent research indicates that CLIP performance can be enhanced through the utilization of visual prompting [24, 26, 31, 32]. Visual prompting (VPT) involves the addition of markers like colorful boxes or circles directly onto an image, aiding in highlighting specific targets in image-language tasks. This technique directs the attention of Vision-Language Models towards desired targets while maintaining the global context. Inspired by this, we also used visual prompting to further evaluate using a separate set of prompts for the lungs. We evaluated three different visual prompts: red bounding box, gray reverse blur, and contour. In our case, we add visual markers on the original image around SAM-generated masks, and pass this set of images to CLIP, as discussed in supplementary Sec. 8. However, for medical datasets, VPT did not perform well. In contrast to such techniques, our method, which involves employing a set of crops of the image according to SAM-generated masks, even while utilizing separate prompts for the left and right lung, still achieves a superior DSC of 0.65 as compared to other prompts as shown in Tab. 3.

Prompt	Box	Reverse blur	Contour	Crops (ours)
DSC	0.49	0.60	0.61	**0.65**

Table 3. Evaluation of visual prompting.

These limitations have offered valuable insights into the failure cases. In the future, we aim to incorporate inference mechanisms to detect such failures and prevent their propagation to the subsequent steps in the pipeline. It will help mitigate the occurrence of such failures and improve performance further.

4.4. Ablations

4.4.1 Different SAM models

In this section, we assess whether employing different variants of SAM can improve the performance. We evaluated all three different versions: ViT-B (base), ViT-L (large), and ViT-H (huge). The results are presented in Tab. 4. Given its superior performance, we opt ViT-H version in our pipeline. Notably, as our approach is training/fine-tuning free and performs test time adaptation for zero-shot segmentation, integrating the ViT-H version imposes no extra training overhead.

Dataset	ViT-B	ViT-L	ViT-H
CC359 [25]	0.80	0.89	**0.94**
X-ray [19]	0.71	0.76	**0.83**
HC18 [28]	0.66	0.76	**0.81**

Table 4. Ablation: Comparison of SAM's variant.

4.4.2 SaLIP vs SAM + CLIP

To assess the performance enhancement brought by our proposed stacking approach of SaLIP, we compared it to SAM + CLIP. Unlike our framework, SAM + CLIP utilizes SAM_{EM} and CLIP exclusively, with the CLIP-retrieved crop considered as the prediction. The results are presented in Tab. 5. Our proposed approach leads to improvement.

4.4.3 Area based filtering

SAM_{EM} employs a grid-wise set of key points to generate masks for each part of the image. The resulting set of masks may include masks for the background or larger regions encompassing the region of interest (ROI). In such

Dataset	SAM-CLIP	SaLIP
CC359 [25]	0.89	**0.94**
X-ray [19]	0.80	**0.83**
HC18 [28]	0.78	**0.81**

Table 5. Ablation: Performance comparison between SAM-CLIP and SaLIP (ours).

cases, CLIP can miss-classify background or larger region encompassing ROI as illustrated in Fig. 6. To tackle this challenge, we implement area-based filtering, determining the optimal area threshold through a random hyperparameter search. To assess the effectiveness of this filtering approach, we compare its results with one obtained by passing all the masks generated by SAM_{EM} to CLIP without any filtering. The comparative results between area filtering and without area filtering are presented in Tab. 6. Our area filtering approach shows an improvement of 3% for brain segmentation [25] and approximately 10% for lung and fetal head segmentation [19].

Dataset	No Filtering	Filtering (Ours)
CC359 [25]	0.91	**0.94**
X-ray [19]	0.75	**0.83**
HC18 [28]	0.71	**0.81**

Table 6. Ablation: Impact of area-based filtering.

5. Conclusion

In this work, we propose a simple and effective unified framework SaLIP, that leverages the zero-shot segmentation and recognition capabilities of SAM and CLIP respectively. By harnessing both the segment everything and promptable segmentation modes from SAM, with CLIP acting as a bridge between them, we demonstrate effective zero-shot organ segmentation. Unlike other SAM-based segmentation methods, SaLIP is training/fine-tuning free and does not rely on domain expertise or annotated data for segmentation. It is fully adapted at test time without the need for pre-training or additional computational overhead. We employ SAM to segment each region within the image, then leverage CLIP to identify the region of interest by using visually descriptive prompts generated from GPT 3.5. Subsequently, we utilize the retrieved region of interest to prompt SAM for organ segmentation. We validate our framework across three diverse medical imaging datasets, demonstrating its robustness. Our work provides an in-depth exploration of SaLIP for zero-shot organ segmentation. In the future, we aim to expand this work to diverse medical imaging datasets and further improve SaLIP, by integrating an inference mechanism to avoid propagation of failures.

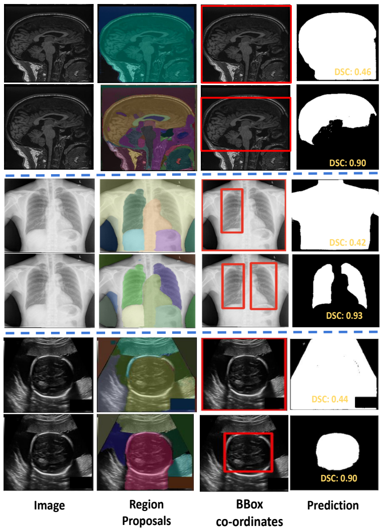


Figure 6. Ablation: the first row for each modality shows results without area filtering, and the second row illustrates the effects of area filtering (ours).

6. Acknowledgment

Dedicated to the memory of ***Dr. Kevin McGuinness***, whose unwavering support and passion for research will forever guide us.

This research is supported by Science Foundation Ireland under grant numbers 18/CRT/6183 (ML-LABS Centre for Research Training) and 12/RC/2289_P2 (Insight Research Centre for Data Analytics).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Risab Biswas. Polyp-sam++: Can a text guided sam perform better for polyp segmentation? *arXiv preprint arXiv:2308.06623*, 2023. 1

[3] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W Remedios, Shunxing Bao, Bennett A Landman, Lee E Wheless, Lori A Coburn, Keith T Wilson, et al. Segment anything model (sam) for digital pathology: Assess zero-shot segmentation on whole slide imaging. *arXiv preprint arXiv:2304.04155*, 2023. 1, 3

[4] 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. 2, 3

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3

[6] Chuanfei Hu and Xinde Li. When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation. *arXiv preprint arXiv:2304.08506*, 2023. 1, 3

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[8] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *Medical Image Analysis*, 92:103061, 2024. 3

[9] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolo, 2023. 1

[10] 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. 1, 2, 3

[11] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21152–21164, 2023. 3

[12] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1

[13] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 3

[14] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 2

[15] Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O'Connor. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 262–271, 2023. 2, 3, 4, 5

[16] Christian Mattjie, Luis Vinicius de Moura, Rafaela Cappelari Ravazio, Lucas Silveira Kupssinskü, Otávio Parraga, Marcelo Mussi Delucis, and Rodrigo Coelho Barros. Exploring the zero-shot capabilities of the segment anything model (sam) in 2d medical imaging: A comprehensive evaluation and practical guideline. *arXiv e-prints*, pages arXiv–2305, 2023. 3

[17] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 1

[18] Sovesh Mohapatra, Advait Gosai, and Gottfried Schlaug. Brain extraction comparing segment anything model (sam) and fsl brain extraction tool. *arXiv preprint arXiv:2304.04738*, 2023. 3

[19] Nikhil Pandey. Chest x-ray masks and labels. `https://www.kaggle.com/datasets/nikhilpandey360/chest-xray-masks-and-labels/data`, 2019. 5, 7, 8, 1, 2

[20] Sumit Pandey, Kuan-Fu Chen, and Erik B Dam. Comprehensive multimodal segmentation in medical imaging: Combining yolov8 with sam and hq-sam models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2592–2598, 2023. 1

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[22] 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. 1, 3

[23] 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*, pages 234–241. Springer International Publishing, 2015. 5

[24] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11987–11997, 2023. 7

[25] Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo. An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, 170:482–494, 2018. 5, 7, 8

[26] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 7

[27] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020. 3

[28] Thomas L A van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2D ultrasound images. *PLoS One*, 13(8):e0200412, 2018. 5, 7, 8, 1, 2

[29] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language (EMNLP)*, pages 3876–3887, 2022. 3

[30] Junde Wu, Rao Fu, Huihui Fang, Yuanpei Liu, Zhaowei Wang, Yanwu Xu, Yueming Jin, and Tal Arbel. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 2

[31] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36, 2024. 7, 2

[32] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38, 2024. 7

[33] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. CXR-CLIP: Toward large scale chest x-ray Language-Image pre-training. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 101–111. Springer Nature Switzerland, 2023. 3

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

[35] Shaoteng Zhang, Jianpeng Zhang, Yutong Xie, and Yong Xia. TPRO: Text-Prompting-Based weakly supervised histopathology tissue segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 109–118. Springer Nature Switzerland, 2023. 3

[36] Yixiao Zhang, Xinyi Li, Huimiao Chen, Yaoyao Yuille, Alan Land Liu, and Zongwei Zhou. Continual learning for abdominal multi-organ and tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 35–45. Springer Nature Switzerland, 2023. 3

[37] Tao Zhou, Yizhe Zhang, Yi Zhou, Ye Wu, and Chen Gong. Can sam segment polyps? *arXiv preprint arXiv:2304.07583*, 2023. 1