



MedCLIP-SAMv2: Towards universal text-driven medical image segmentation

Taha Koleilat ^a,^{*}, Hojat Asgariandehkordi ^a, Hassan Rivaz ^a, Yiming Xiao ^b

^a Department of Electrical and Computer Engineering, Concordia University, Montreal, Quebec, Canada

^b Department of Computer Science and Software Engineering, Concordia University, Montreal, Quebec, Canada

ARTICLE INFO

Keywords:

Text-driven image segmentation
Vision-language models
Foundation models
Weakly supervised segmentation

ABSTRACT

Segmentation of anatomical structures and pathologies in medical images is essential for modern disease diagnosis, clinical research, and treatment planning. While significant advancements have been made in deep learning-based segmentation techniques, many of these methods still suffer from limitations in data efficiency, generalizability, and interactivity. As a result, developing robust segmentation methods that require fewer labeled datasets remains a critical challenge in medical image analysis. Recently, the introduction of foundation models like CLIP and Segment-Anything-Model (SAM), with robust cross-domain representations, has paved the way for interactive and universal image segmentation. However, further exploration of these models for data-efficient segmentation in medical imaging is an active field of research. In this paper, we introduce MedCLIP-SAMv2, a novel framework that integrates the CLIP and SAM models to perform segmentation on clinical scans using text prompts, in both zero-shot and weakly supervised settings. Our approach includes fine-tuning the BiomedCLIP model with a new Decoupled Hard Negative Noise Contrastive Estimation (DHN-NCE) loss, and leveraging the Multi-modal Information Bottleneck (M2IB) to create visual prompts for generating segmentation masks with SAM in the zero-shot setting. We also investigate using zero-shot segmentation labels in a weakly supervised paradigm to enhance segmentation quality further. Extensive validation across four diverse segmentation tasks and medical imaging modalities (breast tumor ultrasound, brain tumor MRI, lung X-ray, and lung CT) demonstrates the high accuracy of our proposed framework. Our code is available at <https://github.com/HealthX-Lab/MedCLIP-SAMv2>.

1. Introduction

With the growing availability of radiological technologies, there is an increasing demand for precise and efficient medical image segmentation to support the study, diagnosis, and treatment of various medical conditions (Siuly and Zhang, 2016). Deep learning (DL) techniques have emerged as state-of-the-art (SOTA) in this field; however, they face three key challenges that hinder their broader clinical adoption. **First**, the scarcity of large, well-annotated datasets presents a major obstacle to DL model development. **Second**, the lack of interactivity and interpretability undermines trust in these methods. **Finally**, most medical DL models are trained for specific tasks and contrasts/modalities, limiting their flexibility. While several self-supervised and weakly supervised approaches (Baevski et al., 2023; Chen et al., 2020; Taleb et al., 2021) have been introduced to improve training efficiency, and explainable AI (XAI) techniques, including uncertainty estimation (Loquercio et al., 2020; Liu et al., 2020) and saliency maps (Arun et al., 2021; Bae et al., 2020) are under active investigation, cross-domain generalization remains a major challenge. Despite these efforts,

existing models often fail to generalize across modalities/tasks and lack the flexibility required for interactive clinical deployment. Addressing these limitations requires a framework that can operate in a zero-shot setting, adapt to multiple imaging types/tasks, and incorporate human guidance.

Recently, the introduction of foundation models, such as Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021) and Segment Anything Model (SAM) (Kirillov et al., 2023) has paved the way for interactive and universal medical image segmentation. Several research groups have adapted CLIP and SAM for radiological tasks, including the development of BiomedCLIP (Zhang et al., 2023) and MedSAM (Ma et al., 2024), which were pre-trained on vast amounts of biomedical data. These models promise cross-modal alignment and segmentation flexibility, but unlocking their full potential for clinical imaging tasks requires further adaptation and fine-tuning. Particularly, to mitigate the dependence on precise drawing-based visual prompts from trained eyes for SAM models, CLIP could offer an alternative

* Corresponding author.

E-mail address: taha.koleilat@mail.concordia.ca (T. Koleilat).

<https://doi.org/10.1016/j.media.2025.103749>

Received 18 February 2025; Received in revised form 30 May 2025; Accepted 22 July 2025

Available online 5 August 2025

1361-8415/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

mechanism to provide these visual prompts through interfacing the user with natural languages, which are more flexible, intuitive, and scalable.

Although CLIP training primarily operates at a global level for image-text mapping, research (Fu et al., 2024) has revealed that these models can encode rich feature representations of images. This allows us to establish the relationship between global textual information and local visual features (Zhou et al., 2022; Rao et al., 2022), which can be exploited for efficient zero-shot medical image segmentation, enabling broader applicability even in data-scarce settings, as we explored for the first time in our MICCAI 2024 paper (Koleilat et al., 2024b). However, transferring CLIP's success from natural images to radiology is non-trivial due to the complexity of medical descriptions and the subtlety of diagnostic imaging features. While adapting CLIP to the medical image domain may seem attractive, it is non-trivial and requires substantial ground truth labels to fine-tune the model effectively, especially for segmentation downstream tasks (Poudel et al., 2024). The lack of large, high-quality annotated datasets in medical imaging further exacerbates this challenge. This motivates the use of biomedical domain-specific variants such as BiomedCLIP (Zhang et al., 2023), which are better suited to capture disease-relevant radiological representations, as well as effective fine-tuning loss functions to establish more effective cross-modal learning in radiological applications, such as pathology localization, segmentation, and diagnosis.

On the other hand, as interest in SAM grows, to mitigate its reliance on visual prompts (e.g., point and/or bounding box) for segmentation, which require prior clinical knowledge, recent methods have emerged to fine-tune SAM without these prompts (Chen et al., 2024; Hu et al., 2023), generate prompts via Class Activation Maps (CAM) from classification tasks (Li et al., 2025, 2023b; Liu and Huang, 2024), and refine its output using weak supervision (Yang and Gong, 2024; Chen et al., 2023; Huang et al., 2023). These approaches point to a growing interest and a new avenue for frameworks that can bridge vision-language understanding with interactive segmentation, instead of relying on manual drawing-based prompts.

Recently, to address these challenges, we introduced MedCLIP-SAM in MICCAI2024 (Koleilat et al., 2024b), which leverages BiomedCLIP (Zhang et al., 2023) to generate text-based box prompts for SAM (Kirillov et al., 2023) towards interactive and universal medical image segmentation, in both zero-shot and weakly supervised settings. Following the preliminary success, further improvement and exploration of the framework are necessary to elevate performance and gain deeper insights into the CLIP and SAM foundation models for medical imaging applications. Therefore, in this paper, we propose **MedCLIP-SAMv2**, a significantly enhanced framework that builds on MedCLIP-SAM by refining CLIP-SAM integration, saliency-guided prompt generation, and uncertainty-aware weak supervision to better harness the synergy of these foundation models for universal, scalable medical segmentation.

Specifically, the prominent upgrades for the newly proposed MedCLIP-SAMv2 framework from the original method include:

- We investigated different saliency map generation techniques for CLIP models, where we replaced gScoreCAM (Chen et al., 2022) with M2IB (Wang et al., 2024), which, when combined with our fine-tuning of BiomedCLIP (Zhang et al., 2023), significantly improved zero-shot segmentation accuracy.
- We enhanced weakly supervised segmentation results and interpretability from the previous framework by training nnUNet (Isensee et al., 2021) using pseudo-labels while providing uncertainty estimation via checkpoint ensembling (Zhao et al., 2022).
- The validation was expanded by incorporating an additional Lung CT dataset, thereby covering four key radiological modalities: CT, MRI, ultrasound, and X-ray. This comprehensive testing further demonstrates the framework's versatility and robustness across diverse segmentation tasks.
- We investigated and optimized advanced text prompt engineering strategies by leveraging large language model (LLM) reasoning and various ensembling methods, which are shown to significantly boost zero-shot segmentation performance.
- Significantly more extensive experiments were conducted for further validation of our framework's design components, including testing different SAM backbones and visual prompt types. We meticulously evaluate the necessity of each component in our framework and demonstrate their individual contribution to the overall performance enhancement.

The proposed MedCLIP-SAMv2 framework addresses fundamental limitations in prior works by combining cross-modal alignment from BiomedCLIP with SAM's flexible segmentation capabilities, thereby advancing towards practical, universal, and interactive segmentation tools in radiology. The newly proposed MedCLIP-SAMv2 framework is more accurate, advancing further towards universal text-driven medical image segmentation with an increase of 13.07% and 11.21% in Dice score for zero-shot and weakly supervised paradigms, respectively. Our main contributions are threefold: **First**, we propose a new CLIP training/fine-tuning loss function called Decoupled Hard Negative Noise Contrastive Estimation (DHN-NCE). **Second**, we introduce a text-driven zero-shot medical segmentation method, combining CLIP and SAM for radiological tasks. **Lastly**, we explore a weakly-supervised strategy to further refine zero-shot segmentation results with uncertainty estimation. Our proposed framework is extensively validated across four distinct segmentation tasks and modalities, including breast tumor segmentation in ultrasound, brain tumor segmentation in MRI, and lung segmentation in chest X-ray and CT.

2. Related work

The core challenge in medical image segmentation lies in generalizing to new tasks with minimal or no annotated training data. Conventional fully supervised segmentation requires large, curated datasets, often impractical in medical imaging due to cost and expertise barriers. As a result, various alternatives have emerged: (1) few-shot segmentation, where limited support examples are provided; (2) weakly supervised methods with coarse annotations; and (3) zero-shot or test-time adaptation strategies that leverage pretrained models and prompts. Our work falls in the third category, proposing a universal segmentation method that eliminates the need for retraining or annotations by using BiomedCLIP and SAM through multimodal prompt ensembling.

2.1. CLIP in medical domain

Several works have utilized CLIP for medical images and texts. Despite being trained on 400 million natural image-text pairs, CLIP's performance suffers on medical tasks. For this reason, works like PubMedCLIP (Eslami et al., 2023) suggested fine-tuning CLIP on a set of PubMed articles for the task of Medical Question-Answering; Zhang et al. (2023) later showed PubMedCLIP's poor performance on cross-modal retrieval tasks (worse than CLIP). On the other hand, MedCLIP (Wang et al., 2022) proposed a technique to utilize decoupled images and texts in training to augment data while Windsor et al. (2023) explored different methods of enhancing the performance of vision-language models for medical domain tasks in a limited data setting. Alternatively, Wu et al. (2023a) proposed a method of enhancing the text in medical reports by simplifying the sentence complexity. Moreover, other works like (Keicher et al., 2023; Tiu et al., 2022) have utilized CLIP for the task of pathology detection and medical report generation. However, notably, almost all mentioned works (Wang et al., 2022; Windsor et al., 2023; Wu et al., 2023a; Keicher et al., 2023; Tiu et al., 2022) only utilized Chest X-ray data for their proposed methods. BiomedCLIP (Zhang et al., 2023) is by far the most recent work for multi-modal medical data on a large scale, which was shown to be

superior for cross-modal retrieval accuracy. Notable studies (Koleilat et al., 2025; Poudel et al., 2024) have investigated the transfer capabilities of BiomedCLIP in downstream tasks such as classification and segmentation. However, its adaptability remains largely unexplored compared to the extensive body of CLIP literature. To the best of our knowledge, our work is the first to explore the potential of BiomedCLIP in zero-shot segmentation tasks, paving the way for more efficient usage in medical imaging where annotations are scarce.

2.2. SAM for medical imaging segmentation

With the advent of SAM, a foundation model for image segmentation that enables zero-shot generalization through a promptable architecture consisting of a powerful image encoder, a flexible prompt encoder, and a lightweight mask decoder, a myriad of research has been dedicated to adapting it for medical imaging purposes. MedSAM (Ma et al., 2024) provided a large-scale fine-tuning of SAM on about 1 million medical image-mask pairs and demonstrated superior performance when it comes to multiple segmentation tasks. AutoSAM (Shaharabany et al., 2023) offered a more efficient approach to fine-tuning SAM on medical images through training the prompt encoder and developing a lightweight deconvolution mask decoder for medical segmentation tasks. Cheng et al. (2023) found that bounding boxes gave the best results when prompting SAM across 12 different medical segmentation tasks, and Huang et al. (2023) proposed a pseudo-mask correction framework to enhance noisy labels generated from SAM for medical images that can be used for further fine-tuning. Finally, Gong et al. (2023) replaced SAM’s mask decoder with a 3D convolutional neural network so that volumetric medical images can be supported. While these approaches adapt SAM for medical imaging, they typically rely on retraining, prompt augmentation, or user input. Our approach avoids all supervision by using text prompts derived from BiomedCLIP, with no point, box, or mask annotations required during inference.

2.3. Few-shot, weakly supervised, and test-time adaptation approaches

Recent works have extended few-shot segmentation to the medical domain by integrating prompt-based methods and support-query frameworks. UniverSeg (Butoi et al., 2023) trains a shared universal segmentation model that generalizes across medical tasks using a task-agnostic architecture. ProtoSAM (Ayzenberg et al., 2024) combines SAM with prototypical learning to leverage a few annotated examples per task. Similarly, Self-Prompt-SAM (Wu et al., 2023b) generates pseudo-prompts from support images for few-shot segmentation without heavy retraining. While effective, these methods still rely on a few support examples per new task and often require task-specific adaptation. In contrast, our approach performs segmentation purely from text prompts, requiring no support images or finetuning.

Self-configuring models, such as nnUNet (Isensee et al., 2021) serve as a robust, fully supervised baseline but require dense annotations. Our weakly supervised nnUNet variant operates with pseudo-masks derived via multimodal saliency and ensemble voting, showing that large vision-language models can bootstrap segmentation performance without needing precise pixel-level supervision. On the other hand, SaLIP (Aleem et al., 2024) proposed a test-time adaptation method for segmenting new classes in medical images by optimizing prompts during inference. Similarly, SAMAug (Dai et al., 2024) augments prompts for SAM to enhance point-level segmentation in unseen domains. While effective, these methods involve adaptation at inference time, increasing computational cost and complexity. Our method avoids this by using static ensembling of pre-generated text prompts, requiring no tuning or optimization at test time.

Overall, our method contributes to the growing body of foundation model-driven segmentation by providing a modular, label-free, and inference-efficient pipeline that bridges CLIP-based text semantics with SAM’s universal segmentation capabilities.

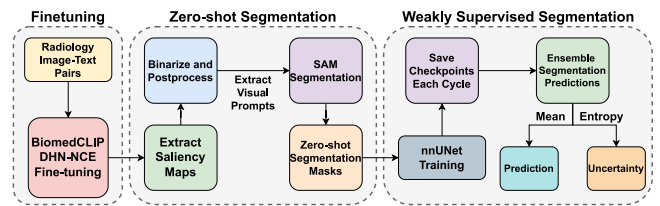


Fig. 1. A general overview of the essential components.

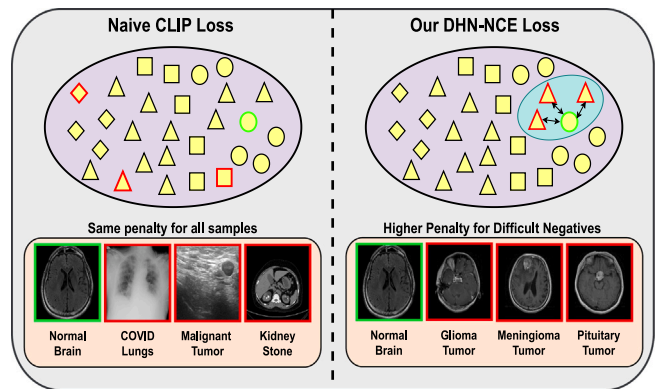


Fig. 2. Comparison of the standard CLIP loss, which applies uniform penalties to all examples regardless of difficulty, with our DHN-NCE loss, which prioritizes harder examples. The DHN-NCE loss enhances the differentiation of medical cases by appropriately penalizing close negatives through adaptive weighting formulas. Green outline represents the anchor example while the red outline represents the negative examples.

3. Methods

A full overview of the proposed MedCLIP-SAMv2 framework is presented in Fig. 3, which is organized into three distinct stages: (1) BiomedCLIP fine-tuning employing our new DHN-NCE loss, (2) zero-shot segmentation guided by text-prompts, and (3) weakly supervised segmentation for potential label refinement. We additionally showcase a summary of the main components of the framework in Fig. 1 for the readers’ easy reference.

3.1. Efficient DHN-NCE fine-tuning

CLIP-like models are commonly trained on extensive datasets consisting of images paired with their corresponding textual descriptions. These models employ an image encoder and a text encoder to extract features, representing them as vectors in a shared dimensional space¹: $\mathbf{I}_{p,i}$ for images and $\mathbf{T}_{p,i}$ for texts. Through the mechanism of contrastive learning, CLIP aligns semantically related image-text pairs by minimizing their distance in the embedding space while maximizing the separation of unrelated pairs. This shared embedding framework facilitates a cohesive understanding of multimodal data. Although BiomedCLIP (Zhang et al., 2023) was trained on medical charts/images and clinical texts, further fine-tuning can significantly enhance its performance on tasks specific to medical imaging. In traditional CLIP training with the InfoNCE loss (Oord et al., 2018), the *negative-positive-coupling* (NPC) effect (Yeh et al., 2022) can reduce learning efficiency, especially with smaller batch sizes. Additionally, for medical images, distinguishing between subtle differences in cases within the same imaging category can be challenging. To address these issues,

¹ It is important to note that CLIP utilizes the global [CLS] tokens of the final vision and text encoder layers before projection to the shared embedding space.

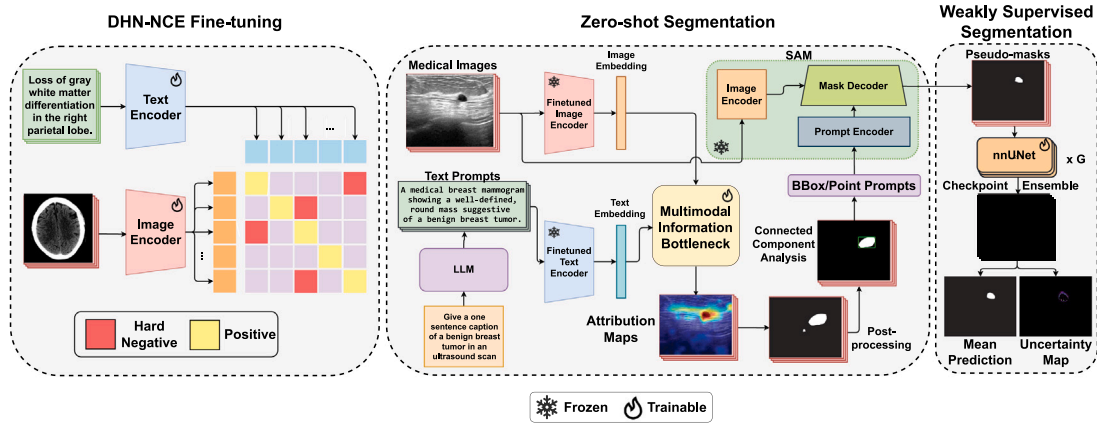


Fig. 3. An overview of the proposed MedCLIP-SAMv2 framework.

we propose the Decoupled Hard Negative Noise Contrastive Estimation (DHN-NCE) loss, which (1) combines the InfoNCE loss (Oord et al., 2018) with hard negative sampling (Robinson et al., 2021), emphasizing “close samples”, and (2) incorporates decoupled contrastive learning (Yeh et al., 2022) by removing the positive term in the denominator, allowing for smaller batch sizes.

Original InfoNCE Loss: The standard InfoNCE loss for contrastive learning is formulated as follows:

$$\mathcal{L}_{\text{InfoNCE}} = - \sum_{i=1}^B \log \frac{\exp(\mathbf{z}_i^T \mathbf{z}_i^+ / \tau)}{\sum_{j=1}^B \exp(\mathbf{z}_i^T \mathbf{z}_j / \tau)} \quad (1)$$

where B is the batch size, \mathbf{z}_i represents the feature embedding of the anchor sample, \mathbf{z}_i^+ is the positive pair for \mathbf{z}_i , τ is the temperature parameter, and B is the batch size.

InfoNCE for Vision-Language Matching: To get a vision-language contrastive loss, we replace generic embeddings with image ($\mathbf{I}_{p,i}$) and text ($\mathbf{T}_{p,i}$) embeddings. In this context, $t \rightarrow v$ refers to text-to-image, while $v \rightarrow t$ indicates image-to-text:

$$\mathcal{L}^{v \rightarrow t} = - \sum_{i=1}^B \log \frac{\exp(\mathbf{I}_{p,i}^T \mathbf{T}_{p,i} / \tau)}{\sum_{j=1}^B \exp(\mathbf{I}_{p,i}^T \mathbf{T}_{p,j} / \tau)} \quad (2)$$

$$\mathcal{L}^{t \rightarrow v} = - \sum_{i=1}^B \log \frac{\exp(\mathbf{T}_{p,i}^T \mathbf{I}_{p,i} / \tau)}{\sum_{j=1}^B \exp(\mathbf{T}_{p,i}^T \mathbf{I}_{p,j} / \tau)} \quad (3)$$

Decoupling Positives and Negatives: Expanding the logarithm and separating terms of Eq. (2) and (3), we obtain:

$$\mathcal{L}^{v \rightarrow t} = - \sum_{i=1}^B \left[\frac{\mathbf{I}_{p,i}^T \mathbf{T}_{p,i}}{\tau} - \log \sum_{j=1}^B \exp(\mathbf{I}_{p,i}^T \mathbf{T}_{p,j} / \tau) \right] \quad (4)$$

$$\mathcal{L}^{t \rightarrow v} = - \sum_{i=1}^B \left[\frac{\mathbf{T}_{p,i}^T \mathbf{I}_{p,i}}{\tau} - \log \sum_{j=1}^B \exp(\mathbf{T}_{p,i}^T \mathbf{I}_{p,j} / \tau) \right] \quad (5)$$

Since the summation in the denominators of Eq. (2) and (3) includes both the positive pair ($j = i$) and the negatives ($j \neq i$), we split it as:

$$\sum_{j=1}^B \exp(\mathbf{I}_{p,i}^T \mathbf{T}_{p,j} / \tau) = \exp(\mathbf{I}_{p,i}^T \mathbf{T}_{p,i} / \tau) + \sum_{j \neq i} \exp(\mathbf{I}_{p,i}^T \mathbf{T}_{p,j} / \tau) \quad (6)$$

Following the positive-negative decoupling approach in Yeh et al. (2022), we remove the positive pair and obtain the decoupled vision-language contrastive loss:

$$\mathcal{L}^{v \rightarrow t} = - \sum_{i=1}^B \frac{\mathbf{I}_{p,i}^T \mathbf{T}_{p,i}}{\tau} + \sum_{i=1}^B \log \left(\sum_{j \neq i} e^{\mathbf{I}_{p,i}^T \mathbf{T}_{p,j} / \tau} \right) \quad (7)$$

$$\mathcal{L}^{t \rightarrow v} = - \sum_{i=1}^B \frac{\mathbf{T}_{p,i}^T \mathbf{I}_{p,i}}{\tau} + \sum_{i=1}^B \log \left(\sum_{j \neq i} e^{\mathbf{T}_{p,i}^T \mathbf{I}_{p,j} / \tau} \right) \quad (8)$$

Applying Hardness Weights: The resulting DHN-NCE loss function, $\mathcal{L}_{\text{DHN-NCE}}$, employs weighting functions ($\mathcal{W}_{\mathbf{I}_{p,i} \mathbf{T}_{p,j}}^{v \rightarrow t}$, $\mathcal{W}_{\mathbf{T}_{p,i} \mathbf{I}_{p,j}}^{t \rightarrow v}$) to increase the penalty for negative samples that are close to the anchor, using image-to-text and text-to-image hardness parameters $\beta_1, \beta_2 \geq 0$.

$$\mathcal{L}^{v \rightarrow t} = - \sum_{i=1}^B \frac{\mathbf{I}_{p,i}^T \mathbf{T}_{p,i}}{\tau} + \sum_{i=1}^B \log \left(\sum_{j \neq i} e^{\mathbf{I}_{p,i}^T \mathbf{T}_{p,j} / \tau} \mathcal{W}_{\mathbf{I}_{p,i} \mathbf{T}_{p,j}}^{v \rightarrow t} \right) \quad (9)$$

$$\mathcal{L}^{t \rightarrow v} = - \sum_{i=1}^B \frac{\mathbf{T}_{p,i}^T \mathbf{I}_{p,i}}{\tau} + \sum_{i=1}^B \log \left(\sum_{j \neq i} e^{\mathbf{T}_{p,i}^T \mathbf{I}_{p,j} / \tau} \mathcal{W}_{\mathbf{T}_{p,i} \mathbf{I}_{p,j}}^{t \rightarrow v} \right) \quad (10)$$

$$\mathcal{L}_{\text{DHN-NCE}} = \mathcal{L}^{v \rightarrow t} + \mathcal{L}^{t \rightarrow v} \quad (11)$$

where the hardness weighting formulas are as follows:

$$\mathcal{W}_{\mathbf{I}_{p,i} \mathbf{T}_{p,j}}^{v \rightarrow t} = (B-1) \times \frac{e^{\beta_1 \mathbf{I}_{p,i}^T \mathbf{T}_{p,j} / \tau}}{\sum_{k \neq i} e^{\beta_1 \mathbf{I}_{p,i}^T \mathbf{T}_{p,k} / \tau}} \quad (12)$$

$$\mathcal{W}_{\mathbf{T}_{p,i} \mathbf{I}_{p,j}}^{t \rightarrow v} = (B-1) \times \frac{e^{\beta_2 \mathbf{T}_{p,i}^T \mathbf{I}_{p,j} / \tau}}{\sum_{k \neq i} e^{\beta_2 \mathbf{T}_{p,i}^T \mathbf{I}_{p,k} / \tau}} \quad (13)$$

The weighting functions leverage exponential scaling to amplify the contributions of hard negatives – those with higher similarity scores – while suppressing easier negatives, ensuring the total weight distribution prioritizes these challenging cases (see Fig. 2). By decoupling the positive term from the denominator, DHN-NCE prevents easy positives from diminishing the gradients associated with hard negatives. This mechanism sharpens the model’s focus on refining distinctions for harder cases, enabling more efficient training even with small batch sizes. Such properties make DHN-NCE particularly suited for medical imaging tasks with limited data availability and subtle feature variations.

3.2. Zero-shot medical image segmentation

In this stage, we utilize the fine-tuned BiomedCLIP with the updated parameters $\theta = \{\theta_{\text{img}}, \theta_{\text{text}}\}$ as the backbone model for feature extraction from both images and text prompts. The core segmentation process relies on the Multi-modal Information Bottleneck (M2IB) technique (Wang et al., 2024), which generates visual saliency maps of the target regions by associating text prompts with image regions.

The zero-shot segmentation pipeline can be described as follows:

Image and Text Embedding Extraction: Given input medical images \mathbf{I} and their corresponding text prompts \mathbf{T} , the image encoder Φ_{img} and the text encoder Φ_{text} from the fine-tuned BiomedCLIP model are used to extract embeddings:

$$\mathbf{Z}_{\text{img}} = \Phi_{\text{img}}(\mathbf{I}; \theta_{\text{img}}) \quad (14)$$

$$\mathbf{Z}_{\text{text}} = \Phi_{\text{text}}(\mathbf{T}; \theta_{\text{text}}) \quad (15)$$

LLM Prompt Generation: Accurate and context-aware biomedical text prompts are crucial for effectively guiding vision-language models in medical image understanding. However, crafting such prompts manually often requires substantial clinical expertise and domain knowledge, and thus can be laborious, particularly for non-experts. Moreover, handcrafted prompts can suffer from inconsistency and ambiguity with non-experts, potentially limiting their effectiveness. To address this, we propose to leverage LLMs to automate the generation of semantically rich prompts that can better inform and guide VLMs during inference in the absence of experts. Recent studies, such as LLaVA-Med (Li et al., 2023a), have demonstrated the promise of integrating LLM-generated prompts into the training pipelines of biomedical VLMs, achieving notable improvements in visual grounding and interpretation. Building on this insight, and in light of recent advancements in LLMs like GPT-4, which now exhibit near-human-level accuracy and stability in generating diagnostic medical reports (Krishna et al., 2024; Gertz et al., 2024), we explore how general-purpose LLMs can complement biomedical-specific VLMs (e.g., BiomedCLIP) to enhance performance in downstream tasks. While BiomedCLIP uses PubMed captions for pretraining to align image and text representations, we extend this paradigm by querying GPT-4 to generate task-specific prompts at inference time. For instance, we use the following format to guide attention to salient image regions:

Give a sentence caption that describes unique visual features of [TARGET] in [MODALITY]

We can engineer different prompt configurations ranging from generic to class-specific context captions and we study the effect of these different styles in Section 4.3.1.

Saliency Map Generation: The embeddings \mathbf{Z}_{img} and \mathbf{Z}_{text} are then passed through the Multi-modal Information Bottleneck (M2IB) module (Wang et al., 2024), which learns to align the image and text modalities by maximizing the mutual information between the input image \mathbf{I} and a good representative text prompt \mathbf{T} while filtering out irrelevant information between the image embedding and the input image. By doing so, the process bridges the semantic gap between encoded visual and textual features ensuring that embeddings emphasize the content that is jointly relevant across image and text. Specifically, the M2IB module introduces a stochastic information bottleneck $\lambda_S \in \mathbb{R}^{H \times W}$ such that $0 \leq \lambda_S \leq 1$ where H, W are the respective height and width of the input image \mathbf{I} . This produces a continuous visual saliency map for the image representing the importance of each pixel concerning the text prompt. This visual saliency map is produced by optimizing the following objective function:

$$\lambda_S = MI(\mathbf{Z}_{\text{img}}, \mathbf{Z}_{\text{text}}; \theta) - \gamma \times MI(\mathbf{Z}_{\text{img}}, \mathbf{I}; \theta) \quad (16)$$

where MI is the mutual information operation and γ is a hyperparameter that balances the trade-off between relevance and compression.

Post-processing for Initial Segmentation: To obtain a discrete pixel-wise segmentation, we apply Otsu's thresholding (Otsu, 1979) to the saliency map λ_S , automatically determining an optimal threshold η^* that separates foreground (regions of interest) from background by minimizing intra-class variance. The binarized segmentation is then given by:

$$\mathbf{Y}_{\text{otsu}} = \begin{cases} 1, & \lambda_S(x, y) \geq \eta^* \\ 0, & \lambda_S(x, y) < \eta^* \end{cases} \quad (17)$$

After thresholding, small, disconnected contours may still exist. To refine the segmentation and ensure robust results, we perform Connected Component Analysis on the identified contours C . For each connected component $c \in C$, we compute a confidence score based on the saliency map λ_S . The confidence of a connected component is derived as follows:

$$\text{Confidence}(c) = \frac{\sum_{i \in c} p_i \cdot \hat{y}_i}{\sum_{i \in c} \hat{y}_i}, \quad (18)$$

where p_i is the probability that pixel i belongs to the foreground class derived from the saliency map λ_S , and \hat{y}_i is the predicted binary label for pixel i derived from the thresholded discrete map \mathbf{Y}_{otsu} . Using this confidence score, we identify the most confident connected components to form the final coarse segmentation:

$$\mathbf{Y}_{\text{coarse}} = \{c \in C : \text{Confidence}(c) > \eta_c\}, \quad (19)$$

where η_c is a confidence threshold that was set to 0.5, which we found to be robust across samples, as most low-confidence components corresponded to small, irrelevant blobs or artifacts. This approach retains all high-confidence connected components, allowing the framework to support **multiple** foreground regions within the same class. It enhances the initial segmentation by filtering out unreliable areas and retaining only the most confident and well-defined regions.

Segmentation Refinement via SAM: The initial segmentation is used as input to the Segment Anything Model, which refines the segmentation by taking visual prompts \mathbf{V} (e.g., bounding boxes or points) derived from the post-processed clusters. For bounding boxes, we calculate 4 box coordinates (bounding boxes) that enclose each connected contour in the coarse segmentation, while for points, we randomly sample different points that lie within the contour. The final zero-shot segmentation mask $\mathbf{Y}_{\text{zero-shot}}$ is thus produced as:

$$\mathbf{Y}_{\text{zero-shot}} = \text{SAM}(\mathbf{Y}_{\text{coarse}}; \mathbf{V}) \quad (20)$$

3.3. Uncertainty-aware weakly supervised segmentation

To further enhance the segmentation accuracy, the zero-shot segmentation results $\mathbf{Y}_{\text{zero-shot}}$ are then used as pseudo-labels with the input medical images \mathbf{I} to train a segmentation network \mathbf{M} in a weakly supervised manner. Thus, the training data will be $\mathcal{T} = \{(\mathbf{I}, \mathbf{Y}_{\text{zero-shot}})\}$. Building on the recent work by Zhao et al. (2022), checkpoint ensembling has demonstrated superior effectiveness in uncertainty estimation for medical image segmentation when compared to techniques such as Monte Carlo Dropout and mean-field Bayesian Neural Networks. This finding is particularly relevant in the context of the nnUNet framework (Isensee et al., 2021). Given a total number of epochs E , the training process is divided into D cycles composed of $E_d = \frac{E}{D}$ epochs, and during each cycle, we save G_d checkpoints of the model. Importantly, this checkpoint strategy adds no delays to the training process, as it involves saving alternate checkpoints of the same model rather than training separate models. Unlike the standard nnUNet protocol that trains across cross-validation folds, we trained on the entire training set in a single run and validated using an external validation set. This not only improves computational efficiency, but also enhances robustness when using pseudo-labels, as high-quality labels in the full dataset help offset noisier ones that could otherwise dominate individual folds. We verify this choice in Section 4.3.6. After completing all training cycles, the probabilistic prediction of the final segmentation $\mathbf{Y}_{\text{final}}$ is obtained by averaging the predictions from the $G = D * G_d$ total checkpoints saved during the training process, providing a Monte-Carlo-like approximation:

$$p(\mathbf{Y}_{\text{final}} | \mathbf{X}; \mathcal{T}) \approx \frac{1}{G} \sum_{n=1}^G p(\mathbf{Y}_{\text{final}} | \mathbf{X}; \mathbf{M}_n) \quad (21)$$

where \mathbf{M}_n represents the weights of the model at the n th checkpoint, and \mathbf{X} are unseen testing input images.

Segmentation Uncertainty Estimation: The variation in predictions across different checkpoints also allows for estimating uncertainty in the final segmentation mask. The generated uncertainty map helps pinpoint regions of the medical scan that exhibit high uncertainty in the prediction. Given R classes in the medical image, the uncertainty for each pixel (i, j) can be computed by calculating the entropy as follows:

$$H(\mathbf{Y}_{\text{final},(i,j)}) = - \sum_{r=1}^R h(r) \log h(r) \quad (22)$$

where

$$h(r) = p(\mathbf{Y}_{\text{final},(i,j)} = r|\mathbf{X}; \mathcal{T}) \quad (23)$$

3.4. Datasets and experimental setup

3.4.1. BiomedCLIP fine-tuning

We employed the public MedPix (Siragusa et al., 2024) dataset, which contains various radiological modalities, to fine-tune the BiomedCLIP model (Zhang et al., 2023) with our DHN-NCE loss. The base encoders for images and text were the Vision Transformer (ViT) and PubMedBERT (Zhang et al., 2023), respectively. The MedPix dataset was cleaned by removing special characters, trimming leading and trailing white spaces, and excluding samples with captions shorter than 20 characters. All images were resized to 224×224 pixels and normalized according to the RGB channel means and standard deviations used in the original CLIP model (Radford et al., 2021). We performed an 85%–15% split, resulting in 20,292 training images and 3515 validation images. Fine-tuning was performed with a learning rate of $1\text{E}-6$, a 50% decay rate, and a batch size of 64.

To validate the fine-tuning quality of BiomedCLIP, we assessed the top-1 and top-2 accuracy of matching retrievals for both image-to-text and text-to-image on the ROCO (Radiology Objects in COntext) dataset (Pelka et al., 2018), which contains approximately 7042 multimodal medical images covering a wide range of radiological cases. We ran the experiments five times with a batch size of 50, using shuffling to randomize image-text pairs (resulting in 70,420 shuffled examples). In addition, we compared different SOTA loss functions for fine-tuning, including InfoNCE (Oord et al., 2018), DCL (Yeh et al., 2022) and HN-NCE (Radenovic et al., 2023) against our DHN-NCE loss. For a fair comparison, all strategies were trained using the same hyperparameters ($\tau = 0.6$, learning rate = $1\text{E}-6$), with the hardness parameters for HN-NCE and DHN-NCE set to $\beta_1 = \beta_2 = 0.15$. As a reference, we also included baseline results from pre-trained BiomedCLIP (Zhang et al., 2023), PMC-CLIP (Lin et al., 2023), and CLIP (Radford et al., 2021).

3.4.2. Datasets

To evaluate the zero-shot and weakly supervised segmentation results, as well as various design elements of the proposed MedCLIP-SAMv2 framework, we utilized four public datasets, each representing different radiology modalities and tasks. These datasets, which include segmentations of breast tumors, brain tumors, and lungs, were divided into training, validation, and testing sets as follows:

- **Breast Tumor Ultrasound:** The Breast Ultrasound Images dataset (BUSI) (Al-Dhabyani et al., 2020), containing 600 images of benign and malignant tumors for training. Additionally, 50 and 113 images from the UDIAT dataset (Byra et al., 2020) were used for validation and testing, respectively. The images in this dataset do not have a fixed resolution, meaning each image has its own unique size, which we preserved during processing.
- **Brain Tumor MRI:** The Brain Tumor dataset (Cheng, 2017), comprising 1462 T1-weighted MRI scans for training, 1002 for validation, and 600 for testing. All MRI slices had a resolution of 512×512 pixels, and were preprocessed to discard empty slices as provided by the dataset.
- **Lung Chest X-ray:** The COVID-19 Radiography Database (COVID-QU-Ex) (Chowdhury et al., 2020; Rahman et al., 2021) is divided into 16,280 chest X-rays (normal, lung opacity, viral pneumonia, and COVID-19 cases) for training, 1372 for validation, and 957 for testing. The images had a resolution of 299×299 pixels.
- **Lung CT:** CT scans from Konya (2020), consisting of segmentation masks for fibrotic diseased lungs from 107 patients, split into 7959 slices for training, 3010 for validation, and 1800 for testing. The split was done by patient ID to prevent data leakage. All slices were provided at a resolution of 512×512 pixels, and only non-empty slices were used as preprocessed by the dataset authors.

3.4.3. Experimental setup and metrics

We performed a comprehensive comparison of segmentation quality using the initial labels derived from post-processed M2IB results, zero-shot pseudo-masks, and weakly supervised outputs on the specified testing datasets. To generate visual prompts from BiomedCLIP, all images were resized to 224×224 pixels, as required by the model. For inference with SAM, images were resized to 1024×1024 , in line with SAM’s input specification. All zero-shot segmentation outputs were subsequently **resized back to the original resolution** of the input images for nnUNet training and evaluation. Our zero-shot method was benchmarked against SOTA zero-shot segmentation methods, such as SaLIP (Aleem et al., 2024) and SAMAUG (Dai et al., 2024) and few-shot approaches, such as UniverSeg (Butoi et al., 2023), ProtoSAM (Ayzenberg et al., 2024), and Self-Prompt-SAM (Wu et al., 2023b). Additionally, we compare our weakly supervised method with nnUNet (Isensee et al., 2021) trained on pseudo-labels without checkpoint ensembling. For weakly supervised segmentation, we trained the nnUNet (Isensee et al., 2021) architecture for 600 epochs with 3 cycles for all datasets. The learning rate was initialized to 0.01 and we adopted a cyclical learning rate schedule as described in Zhao et al. (2022), where the learning rate oscillates between a maximum and minimum value throughout each cycle. This allows the model to escape local optima and explore a wider solution space, leading to more diverse and robust predictions. We saved the last 10 checkpoints in each of the 3 cycles, resulting in 30 total model checkpoints. The final segmentation result is averaged from the predictions of these 30 checkpoints and is later thresholded to create a binary mask.

As part of the ablation studies for zero-shot segmentation, we examined: (1) the impact of fine-tuning BiomedCLIP and the choice of explainable AI (XAI) technique for saliency map generation, (2) the influence of different text prompts on overall segmentation performance, (3) the contribution of each model component to the final performance, and (4) the selection of SAM pre-trained models with various visual prompting strategies. These ablation studies were conducted on the test sets of all four datasets mentioned. In all experiments, Dice-Sørensen Coefficient (DSC) and Normalized Surface Distance (NSD) were used as evaluation metrics. Paired-sample t-tests were also conducted to validate the observed trends, with a p -value of less than 0.05 indicating statistical significance.

4. Results

4.1. Comparison with SOTA methods

Table 1 shows a comparison of the proposed MedCLIP-SAMv2 with different SOTA techniques. Compared to the original MedCLIP-SAM, our approach significantly improved the average DSC from **64.54%** to **77.61%** and NSD from **66.10%** to **81.56%** in the zero-shot setting. Similarly, in the weakly supervised scenario, the average DSC increased from **70.90%** to **82.11%** and NSD from **73.77%** to **87.33%**, even surpassing weakly supervised nnUNet trained on pseudo-labels without checkpoint ensembling on average. Overall, our method significantly outperformed all zero-shot and few-shot SOTA methods across various imaging modalities/tasks ($p < 0.05$), except for Lung X-ray. However, the fully supervised methods still offer higher accuracy than those using limited resources.

4.2. Effectiveness of DHN-NCE

The accuracy of cross-modal retrieval (text-to-image and image-to-text) for the ROCO dataset (Pelka et al., 2018) is shown in Table 2 across different losses for fine-tuning BiomedCLIP, with three pre-trained CLIP models as baselines. It can be seen that domain-specific pre-trained models performed better than CLIP, with the larger-scale pretraining offering better retrieval accuracy while the pre-trained BiomedCLIP demonstrated the highest retrieval accuracy among all

Table 1
Comparison of DSC and NSD values (%) with different few-shot and zero-shot medical image segmentation methods (mean_{std}).

Technique	Method	Breast Ultrasound		Brain MRI		Lung X-ray		Lung CT		All	
		DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑
Zero-shot	SaLIP	44.33 _{10.12}	48.62 _{10.25}	47.96 _{9.14}	50.24 _{9.26}	63.14 _{11.34}	66.44 _{11.58}	76.32 _{11.22}	78.46 _{11.35}	57.94 _{10.49}	60.94 _{10.65}
	SAMAug	56.39 _{10.85}	59.23 _{10.92}	45.71 _{10.34}	48.81 _{11.29}	57.18 _{12.12}	60.08 _{12.34}	44.61 _{10.42}	46.48 _{10.57}	50.97 _{10.96}	53.65 _{11.30}
	MedCLIP-SAM	67.82 _{8.26}	69.12 _{9.12}	66.72 _{5.27}	68.01 _{6.16}	64.49 _{9.09}	65.89 _{10.44}	59.14 _{9.52}	60.47 _{9.98}	64.54 _{8.20}	66.10 _{9.08}
	Ours	77.76 _{9.52}	81.11 _{9.89}	76.52 _{7.06}	82.23 _{7.13}	75.79 _{3.44}	80.88 _{3.52}	80.38 _{5.81}	82.03 _{5.94}	77.61 _{6.82}	81.56 _{7.00}
Weakly supervised	nnUNet	73.77 _{14.48}	79.71 _{14.79}	77.16 _{12.17}	85.21 _{12.60}	70.15 _{6.40}	74.10 _{6.59}	82.24 _{5.12}	85.65 _{4.70}	75.83 _{10.31}	81.17 _{10.52}
	MedCLIP-SAM	58.62 _{5.66}	60.94 _{5.87}	58.80 _{8.63}	61.77 _{8.64}	86.07 _{8.61}	88.65 _{8.09}	80.12 _{3.38}	83.73 _{8.29}	70.90 _{7.92}	73.77 _{7.80}
	Ours	78.87 _{12.29}	84.58 _{12.19}	80.03 _{9.91}	88.25 _{10.04}	80.77 _{4.44}	84.53 _{4.51}	88.78 _{4.43}	91.95 _{4.06}	82.11 _{8.49}	87.33 _{8.46}
One-shot	UniverSeg	40.56 _{5.14}	53.25 _{6.22}	23.81 _{5.45}	35.28 _{6.49}	68.15 _{2.21}	80.09 _{2.16}	54.94 _{8.21}	69.62 _{7.59}	46.87 _{5.67}	59.56 _{5.98}
	ProtoSAM	48.44 _{10.93}	50.24 _{10.84}	45.68 _{15.14}	51.69 _{15.65}	80.75 _{1.40}	85.11 _{1.30}	84.50 _{9.94}	87.62 _{9.72}	64.84 _{10.60}	68.67 _{10.71}
Few-shot (K = 4)	UniverSeg	47.56 _{8.57}	54.25 _{8.71}	53.82 _{10.17}	66.40 _{9.96}	79.25 _{2.10}	84.80 _{1.70}	65.68 _{12.02}	70.56 _{11.67}	61.58 _{9.02}	69.00 _{8.86}
	Self-Prompt-SAM	42.04 _{17.19}	44.30 _{17.64}	46.43 _{15.25}	50.29 _{15.83}	67.97 _{2.89}	71.63 _{2.83}	81.50 _{3.84}	83.40 _{3.77}	59.49 _{11.74}	62.41 _{12.08}
Few-shot (K = 16)	UniverSeg	66.36 _{8.57}	72.22 _{8.30}	62.82 _{7.97}	72.76 _{7.94}	83.44 _{1.54}	87.73 _{1.24}	86.49 _{2.49}	89.96 _{1.94}	74.78 _{6.03}	80.67 _{5.86}
	Self-Prompt-SAM	62.36 _{16.38}	66.01 _{16.92}	52.55 _{15.29}	57.07 _{15.93}	82.49 _{2.50}	86.49 _{2.45}	83.66 _{3.90}	85.49 _{3.84}	70.27 _{11.44}	73.77 _{11.84}
Fully supervised	nnUNet	82.47 _{10.49}	88.32 _{10.77}	87.74 _{6.28}	95.10 _{6.28}	98.72 _{0.65}	99.51 _{0.41}	97.10 _{2.74}	99.18 _{2.13}	84.63 _{6.27}	90.42 _{6.33}
	nnUNet ensemble	84.72 _{10.97}	90.85 _{11.26}	88.82 _{5.93}	95.84 _{5.54}	99.14 _{2.50}	99.82 _{1.93}	98.12 _{4.09}	99.65 _{4.03}	85.43 _{6.68}	91.74 _{6.66}

Table 2
Top-K cross-modal retrieval accuracy (mean_{std}) for CLIP models.

Model	Version	image → text (%)		text → image (%)	
		Top-1	Top-2	Top-1	Top-2
CLIP (Radford et al., 2021)	Pre-trained	26.68 _{0.30}	41.80 _{0.19}	26.17 _{0.20}	41.13 _{0.20}
PMC-CLIP (Lin et al., 2023)	Pre-trained	75.47 _{0.37}	87.46 _{0.11}	76.78 _{0.11}	88.35 _{0.19}
BiomedCLIP (Zhang et al., 2023)	Pre-trained	81.83 _{0.20}	92.79 _{0.13}	81.36 _{0.48}	92.27 _{0.14}
	InfoNCE (Oord et al., 2018)	84.21 _{0.35}	94.47 _{0.19}	85.73 _{0.19}	94.99 _{0.16}
	DCL (Yeh et al., 2022)	84.44 _{0.37}	94.68 _{0.19}	85.89 _{0.16}	95.09 _{0.19}
	HN-NCE (Radenovic et al., 2023)	84.33 _{0.35}	94.60 _{0.19}	85.80 _{0.17}	95.10 _{0.19}
	DHN-NCE (ours)	84.70 _{0.33}	94.73 _{0.16}	85.99 _{0.19}	95.17 _{0.19}

Table 3
Effect of different text prompt templates on the segmentation performance (% , mean_{std}).

Prompt	Breast Ultrasound		Brain MRI		Lung X-ray		Lung CT	
	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑
P0	63.79 _{15.12}	67.89 _{15.08}	70.98 _{7.61}	76.42 _{7.63}	75.79 _{3.44}	80.88 _{3.52}	69.89 _{5.14}	71.83 _{4.98}
P1	67.66 _{14.35}	71.56 _{4.78}	37.19 _{10.98}	39.77 _{11.63}	69.72 _{4.65}	73.52 _{4.83}	–	–
P2	69.04 _{12.45}	73.33 _{12.97}	71.18 _{7.16}	77.19 _{7.14}	63.91 _{4.73}	67.63 _{5.13}	80.38 _{5.81}	82.03 _{5.94}
P3	77.76 _{9.52}	81.11 _{9.89}	76.52 _{7.06}	82.23 _{7.13}	63.92 _{4.88}	67.73 _{4.96}	–	–
P4	59.11 _{16.30}	62.61 _{16.96}	70.16 _{7.86}	73.12 _{8.02}	58.56 _{7.76}	62.48 _{7.92}	65.39 _{10.55}	67.31 _{10.83}
P5	57.94 _{16.11}	61.55 _{17.39}	73.52 _{7.08}	78.78 _{7.13}	58.82 _{8.43}	52.62 _{8.55}	–	–
P6	67.65 _{16.54}	71.02 _{16.89}	69.23 _{8.41}	74.32 _{8.59}	68.95 _{4.91}	72.31 _{4.95}	75.84 _{4.88}	77.56 _{4.97}
P7	65.18 _{17.51}	68.75 _{17.93}	69.81 _{7.86}	75.01 _{7.97}	68.44 _{4.63}	72.09 _{4.81}	–	–
P8	66.10 _{16.25}	69.43 _{16.91}	71.03 _{7.70}	76.02 _{7.85}	66.79 _{5.67}	70.52 _{5.91}	74.11 _{6.40}	75.98 _{6.68}
P9	65.32 _{16.12}	68.69 _{16.85}	73.04 _{7.19}	78.27 _{7.26}	65.98 _{9.32}	68.91 _{9.71}	–	–

pre-trained models. Fine-tuning BiomedCLIP further enhanced its performance. Specifically, BiomedCLIP fine-tuned with DHN-NCE reached **84.70%** top-1 and **94.73%** top-2 in image-to-text retrieval, and **85.99%** top-1 and **95.17%** top-2 in text-to-image retrieval, significantly outperforming other loss functions and the baseline models ($p < 0.01$). Additionally, the benefit of fine-tuning BiomedCLIP with our DHN-NCE loss is further validated with improved segmentation quality across different tasks and image modalities in Tables 4 and 5.

4.3. Ablation experiments

4.3.1. Effect of text prompt designs

We conducted a series of experiments to analyze the impact of various text prompt designs on zero-shot segmentation performance. In particular, we compared six different prompt configurations: **P0** and **P1** include the class name of the object to be segmented, while **P2** and **P3** consist of longer, descriptive single prompts. Note that **P0**, **P2**, **P4**, **P6**, and **P8** are generic text prompts, while **P1**, **P3**, **P5**, **P7**, and **P9** are more nuanced with subtypes of the target object of interests. For example, for Breast Ultrasound, **P0** is “breast tumor” while **P1** can

either be “malignant breast tumor” or “benign breast tumor” depending on the tumor class. For **P2**, we used one descriptive sentence, such as “A medical breast mammogram showing a suspicious, irregularly shaped mass suggestive of a breast tumor.” **P3**, on the other hand, includes descriptive text about a specific tumor subtype, like “A medical breast mammogram showing an irregularly shaped, spiculated mass suggestive of a malignant breast tumor.” **P4-P5**, **P6-P7**, and **P8-P9** are similar to **P2** and **P3**, but they use an ensemble approach by averaging the text embeddings of 5, 20, and 50 different prompts respectively. Here, all descriptive clinical prompts are generated using GPT-4 (Achiam et al., 2023). For Lung CT, we evaluated solely on generic prompts as there is only one class available. As shown in Table 3, the choice of text prompt significantly influences segmentation performance. Class-specific prompts (**P3**) generally yielded better results for smaller structures like breast and brain tumors whereas generic prompts (**P0**, **P2**) performed better for larger structures like lungs in X-ray and CT scans, where simpler, more generic descriptions allowed the model to focus on larger areas. Importantly, LLM-generated prompts (**P2**, **P3**) consistently outperformed manually crafted prompts (**P0**, **P1**) across most tasks, underscoring the benefit of using automatically generated, semantically rich descriptions to guide

Table 4
Effect of different components (% , mean_{std}).

Method	DSC↑	NSD↑
1: Saliency maps	46.23 _{8.58}	50.50 _{8.86}
2: + DHN-NCE fine-tuning	49.10 _{8.46}	53.54 _{8.62}
3: + Post-processing	51.62 _{7.57}	55.23 _{7.47}
4: + Connected component analysis	57.89 _{7.87}	61.54 _{8.02}
5: + SAM	77.61 _{6.82}	81.56 _{7.00}
6: + nnUNet ensemble	82.11 _{8.49}	87.33 _{8.46}

the model. On the other hand, increasing the number of prompts from 5 (**P4**, **P5**) to 20 (**P6**, **P7**) led to a noticeable improvement in segmentation performance across all datasets, except brain MRI. However, expanding from 20 to 50 (**P8**, **P9**) prompts resulted in a slight decrease in performance for most tasks. For the case of brain tumors, performance remained relatively stable regardless of prompt count. This suggests that while a moderate number of diverse prompts can enhance the semantic richness of the text embedding and improve alignment with the image features, too many prompts may introduce redundancy or noise, potentially diluting the relevance of the averaged embedding. These results highlight the importance of balancing diversity and specificity in prompt design for optimal performance. The best prompt configuration for each task is used to generate the results presented in [Table 1](#).

4.3.2. Ablation analysis of algorithm components

[Table 4](#) shows the contribution of each component of our framework in improving the average segmentation performance on all datasets. Starting with saliency maps generated using the M2IB, we achieved a baseline DSC of **46.23%** and an NSD of **50.50%**, providing an initial focus on key regions of interest. Fine-tuning BiomedCLIP with the proposed DHN-NCE loss raised the DSC to **49.10%** and the NSD to **53.54%**. Post-processing the saliency maps further enhanced the segmentation quality, allowing the model to better delineate foreground and background areas by refining the initial segmentation boundaries. Incorporating a connected component analysis step greatly impacted the results, increasing the DSC to **57.89%** and the NSD to **61.54%**, as it eliminated small, irrelevant clusters and reduced noise, improving overall precision. With the integration of SAM and the use of visual prompts, such as bounding boxes or points, our zero-shot method yielded a substantial improvement, achieving a DSC of **77.61%** and an NSD of **81.56%**. Finally, weakly supervised training with checkpoint ensembling further refined the segmentation quality by leveraging pseudo-labels generated from the zero-shot method. By using these pseudo-labels to train a segmentation network from scratch, we were able to reach a final DSC of **82.11%** and an NSD of **87.33%**.

4.3.3. Impact of saliency maps generation methods

As shown in [Table 5](#), M2IB achieved the highest performance across all tasks, with an average DSC of **77.61%** and NSD of **81.56%** when using the fine-tuned BiomedCLIP model. In both its pre-trained and fine-tuned forms, M2IB significantly outperformed gScoreCAM and GradCAM ($p < 0.05$). BiomedCLIP fine-tuning improved the scores across all saliency map techniques on average, with the largest gains seen in M2IB, which improved by **3.92%** in DSC and **4.24%** in NSD compared to its pre-trained version.

4.3.4. Comparison of visual prompts for SAM

[Table 6](#) compares different SAM models and visual prompting techniques. For point-based prompting, we sampled between **5 to 10 random positive points** for each task. We exclusively used positive points, as including negative points or increasing the number of points beyond 10 was found to degrade performance. We see that bounding boxes generally provided the best segmentation performance, except in Lung X-rays, where adding point prompts enhanced results. On the other

Table 5

Comparison between different saliency map techniques as well as the pre-trained and fine-tuned BiomedCLIP on the overall performance (% , mean_{std}).

Model	Technique	All	
		DSC ↑	NSD ↑
Pre-trained BiomedCLIP	M2IB	73.69 _{7.58}	77.32 _{7.43}
	gScoreCAM	58.92 _{6.67}	62.19 _{6.02}
	GradCAM	29.21 _{8.74}	31.36 _{8.44}
Fine-tuned BiomedCLIP	M2IB	77.61 _{6.82}	81.56 _{7.00}
	gScoreCAM	60.52 _{6.41}	63.89 _{6.39}
	GradCAM	30.11 _{8.92}	32.61 _{8.83}

hand, point prompts alone performed worse except in certain tasks, such as Lung X-ray (**75.79%** DSC, **80.88%** NSD). Additionally, we tested an ensemble approach, where SAM was run five times with different randomly sampled positive point prompts, and the resulting masks were averaged. While this ensembling marginally improved performance in some cases (e.g., Lung X-ray), it did not lead to consistent gains over a single point prompt run. Given the high computational cost of this method, we conclude that multi-run point-based ensembling offers limited benefit and is not a practical alternative to single-pass prompting.

In addition, the comparison across SAM, MedSAM, and SAM-Med2D reveals that SAM (ViT-H) consistently outperforms the medical SAM variants, despite not being pre-trained on medical imaging data. It achieves strong performance with bounding box prompts across most modalities and tasks, including challenging ones like Lung CT. SAM (ViT-B), while smaller, still performs on par with or better than the ViT-B-based medical SAM models (SAM-Med2D and MedSAM) in many cases.

4.3.5. Effect of ensembling pseudo-masks

In [Table 7](#), we compared two pseudo-mask ensembling strategies using different combinations of text prompts: (1) ensembling coarse segmentation masks from multiple text prompts before SAM refinement (i.e., “Before Refinement” in [Table 7](#)), and (2) ensembling zero-shot SAM outputs generated from different prompts after individual refinement (i.e., “After Refinement”). Prompt settings included **All** (P0–P3, P6–7), **Class-specific** (P1, P3, P7), **Generic** (P0, P2, P6), and **Single** (best-performing prompt per task). For **All**, **Class-specific**, and **Generic**, we generated one pseudo-mask per sub-strategy in the respective setting (e.g., **Generic** produces three masks from P0, P2, and P6), which were then ensembled either before or after SAM refinement. In contrast, the **Single** configuration used only one prompt (one pseudo-mask output). Ensembling SAM mask outputs consistently outperformed ensembling coarse masks before refining across most tasks. However, we do not observe any improvement in merging pseudo masks from different text prompt configurations. We conclude that using a single, well-designed prompt will often yield even better performance, highlighting the importance of prompt quality for guiding zero-shot segmentation.

4.3.6. Effect of nnunet ensembling strategies

To assess whether more selective or weighted ensemble strategies can improve weakly supervised segmentation performance, we conducted an ablation study comparing different checkpoint ensembling techniques within the nnUNet framework on the Breast Ultrasound and Lung CT datasets. Specifically, we examined four weighting schemes ([Huang et al., 2017](#)): **Average**, which applies uniform averaging across all saved checkpoints; **Early**, which assigns linearly decreasing weights to prioritize early-stage checkpoints; **Late**, which applies linearly increasing weights to favor later-stage checkpoints; and **Middle**, which uses a Gaussian-shaped weighting centered around the midpoint of training checkpoint outcomes, emphasizing mid-stage

Table 6
Comparison between different SAM pre-trained models and visual prompting techniques (% , mean_{std}).

Model	Type	Prompts	Breast Ultrasound		Brain MRI		Lung X-ray		Lung CT	
			DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑
SAM	ViT-B	Points	59.86 _{15.46}	62.67 _{16.02}	52.28 _{16.89}	55.67 _{17.78}	64.49 _{9.01}	67.99 _{9.19}	51.98 _{10.50}	54.30 _{10.52}
		BBoxes	74.27 _{12.36}	78.44 _{12.74}	70.81 _{11.43}	76.59 _{11.61}	55.58 _{11.88}	60.63 _{12.27}	63.49 _{11.64}	66.10 _{12.32}
		Points + BBoxes	64.91 _{11.87}	68.50 _{12.13}	61.91 _{10.57}	66.58 _{10.81}	64.86 _{6.95}	68.98 _{7.11}	52.51 _{7.76}	54.32 _{7.72}
SAM	ViT-H	Points	65.56 _{9.89}	68.20 _{9.97}	65.54 _{8.45}	70.73 _{8.32}	75.79 _{3.44}	80.88 _{3.52}	61.49 _{6.25}	63.90 _{6.74}
		BBoxes	77.76 _{9.52}	81.11 _{9.89}	76.52 _{7.06}	82.23 _{7.12}	70.55 _{5.38}	74.12 _{5.77}	80.38 _{5.81}	82.03 _{5.94}
		Points + BBoxes	74.38 _{10.57}	79.60 _{10.62}	75.48 _{8.66}	80.29 _{8.64}	73.30 _{5.94}	79.22 _{6.12}	62.83 _{6.72}	64.57 _{6.99}
		Points ensemble	62.86 _{13.96}	64.84 _{14.44}	61.06 _{13.03}	64.20 _{13.68}	76.83 _{4.68}	82.41 _{5.72}	63.12 _{7.08}	65.73 _{7.45}
SAM-Med2D	ViT-B	Points	73.12 _{9.51}	75.16 _{9.13}	66.78 _{9.97}	70.12 _{9.75}	60.58 _{7.43}	64.42 _{7.73}	65.94 _{7.17}	68.05 _{7.99}
		BBoxes	75.22 _{10.04}	80.03 _{10.94}	55.21 _{9.85}	61.34 _{9.93}	30.18 _{11.15}	36.35 _{11.23}	63.10 _{8.57}	68.59 _{8.48}
		Points + BBoxes	74.83 _{10.78}	79.50 _{10.12}	67.85 _{10.96}	72.04 _{10.45}	37.23 _{8.69}	44.90 _{9.37}	71.22 _{8.09}	78.05 _{8.11}
MedSAM	ViT-B	BBoxes	63.50 _{11.42}	68.11 _{11.25}	67.68 _{12.75}	73.89 _{12.67}	73.03 _{6.03}	76.23 _{6.02}	62.14 _{7.80}	65.00 _{7.11}

Table 7
Zero-shot evaluation of pseudo-mask ensembling techniques using different text prompts.

Technique	Prompts	Breast Ultrasound		Brain MRI		Lung X-ray		Lung CT	
		DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑	DSC ↑	NSD ↑
Before refinement	All	57.43	61.57	60.00	64.23	54.78	58.65	59.84	61.40
	Class-specific	60.65	64.74	61.78	66.13	54.87	58.91	–	–
	Generic	52.30	56.41	63.17	67.77	55.84	59.76	59.84	61.40
	Single	77.76	81.11	76.52	82.23	75.79	80.88	80.38	82.03
After refinement	All	74.30	78.29	61.49	65.80	68.37	72.63	68.03	69.96
	Class-specific	73.55	77.56	61.83	66.21	65.23	69.56	–	–
	Generic	66.97	70.73	63.47	68.10	67.83	72.09	68.03	69.96
	Single	77.76	81.11	76.52	82.23	75.79	80.88	80.38	82.03

Table 8
Comparison between different nnUNet ensembling techniques on weakly supervised segmentation.

Technique	Weighting	Breast Ultrasound		Lung CT	
		DSC ↑	NSD ↑	DSC ↑	NSD ↑
Checkpoint ensembling	Average	78.87	84.58	88.78	91.95
	Early	77.34	82.92	88.74	91.92
	Middle	77.43	83.00	88.71	91.88
	Late	77.69	83.24	88.75	91.93
2-fold	Average	74.46	80.14	66.51	70.76
5-fold	Average	77.34	83.04	57.08	60.99

checkpoints believed to strike a balance between underfitting and overfitting.

As shown in Table 8, the simple averaging method outperformed all alternative weighting strategies across the two datasets. While late checkpoints intuitively may capture more refined representations, we did not observe any statistically significant improvements compared to uniform averaging. This indicates that the diversity among checkpoints, when averaged uniformly, serves as a strong regularizer and leads to more robust predictions.

We also compared this approach with 2-fold and 5-fold averaging schemes (i.e., ensembles trained on subset splits of the data), which resulted in lower performance, likely due to reduced data exposure per fold and increased sensitivity to pseudo-label noise. These results suggest that while selective weighting is a plausible direction, in practice, simple averaging remains the most reliable and effective ensembling strategy in the weakly supervised setup.

4.4. Evaluation on challenging tasks

Lung nodule segmentation in CT scans is a particularly demanding task due to the high variability in nodule size, irregular shapes, and

the often subtle contrast between nodules and surrounding anatomical structures. This challenge is especially pronounced with small nodules, which can be easily overlooked or confused with nearby tissues, making accurate segmentation difficult even for expert clinicians. We evaluate our method on the UniToChest dataset (Chaudhry et al., 2022), which comprises CT scans from 623 patients with 10,071 lesions annotated for segmentation. The dataset categorizes nodules by size into *Lung Nodules Big* (≥ 10 mm) and *Lung Nodules Small* (< 10 mm), with 11,466/1132/1514 and 7089/580/953 samples allocated to training, validation, and test splits, respectively. All scans are preprocessed and uniformly sliced into 2D images for consistent evaluation. As shown in Table 9, our method consistently outperforms all existing state-of-the-art zero-shot, few-shot, and weakly supervised segmentation techniques across both large and small nodule categories. Notably, our model achieves the highest DSC and NSD scores despite not relying on any additional supervision, highlighting its strong ability to generalize across complex and heterogeneous cases. Despite this relatively strong performance, it still lags behind fully supervised models, particularly in small nodules segmentation. This gap highlights the general advantage of dense annotations in maximizing segmentation accuracy. Nonetheless, our approach offers a promising trade-off between performance and annotation efficiency, and with minimal supervision, it has the potential to close the gap with fully supervised methods, making it a viable and scalable solution for practical clinical deployment.

4.5. Qualitative segmentation results

Lastly, we present qualitative segmentation results across the four imaging modalities evaluated for our proposed method in Fig. 4. Our proposed MedCLIP-SAMv2 consistently produced high-quality segmentation masks in weakly supervised settings. For all datasets except Brain MRI, the initial coarse segmentation was suboptimal. However, it provided a sufficient starting point for the zero-shot approach to refine coarse activation maps. For breast and brain tumors, the zero-shot results were notably better than those for Lung CT and Lung

Table 9
Comparison of DSC and NSD values (%) on Small and Big Lung Nodules (mean_{std}).

Technique	Method	Lung Nodules Big		Lung Nodules Small	
		DSC \uparrow	NSD \uparrow	DSC \uparrow	NSD \uparrow
Zero-shot	SaLIP	13.74 _{11.82}	17.26 _{12.37}	3.42 _{11.91}	4.88 _{12.15}
	SAMAug	16.43 _{12.18}	19.52 _{13.02}	2.91 _{11.66}	3.76 _{11.98}
	MedCLIP-SAM	25.61 _{13.11}	26.83 _{13.49}	4.13 _{12.04}	5.01 _{12.57}
	Ours	40.76 _{13.73}	46.44 _{13.98}	7.93 _{12.28}	10.84 _{14.67}
Weakly supervised	nnUNet	58.27 _{11.76}	66.13 _{12.12}	21.19 _{12.13}	28.31 _{13.01}
	MedCLIP-SAM	29.42 _{12.83}	33.17 _{12.41}	9.64 _{10.97}	10.12 _{11.36}
	Ours	60.64 _{11.02}	68.79 _{11.45}	24.73 _{10.05}	30.77 _{12.57}
One-shot	UniverSeg	3.16 _{4.25}	4.84 _{5.25}	1.41 _{1.67}	2.60 _{2.63}
	ProtoSAM	30.67 _{14.52}	35.10 _{15.34}	3.27 _{6.75}	4.12 _{7.54}
Few-shot (K = 4)	UniverSeg	8.85 _{8.75}	11.98 _{10.60}	1.51 _{3.07}	2.80 _{4.88}
	Self-Prompt-SAM	6.05 _{5.88}	6.86 _{6.01}	1.04 _{4.88}	1.97 _{5.75}
Few-shot (K = 16)	UniverSeg	24.37 _{14.06}	28.61 _{15.58}	3.23 _{3.71}	5.49 _{5.59}
	Self-Prompt-SAM	13.70 _{10.53}	14.05 _{15.21}	2.12 _{4.83}	3.19 _{5.67}
Fully supervised	nnUNet	79.41 _{9.66}	88.82 _{9.76}	50.44 _{14.62}	58.63 _{17.39}
	nnUNet ensemble	81.78 _{8.22}	90.93 _{8.27}	52.03 _{13.74}	60.65 _{16.11}

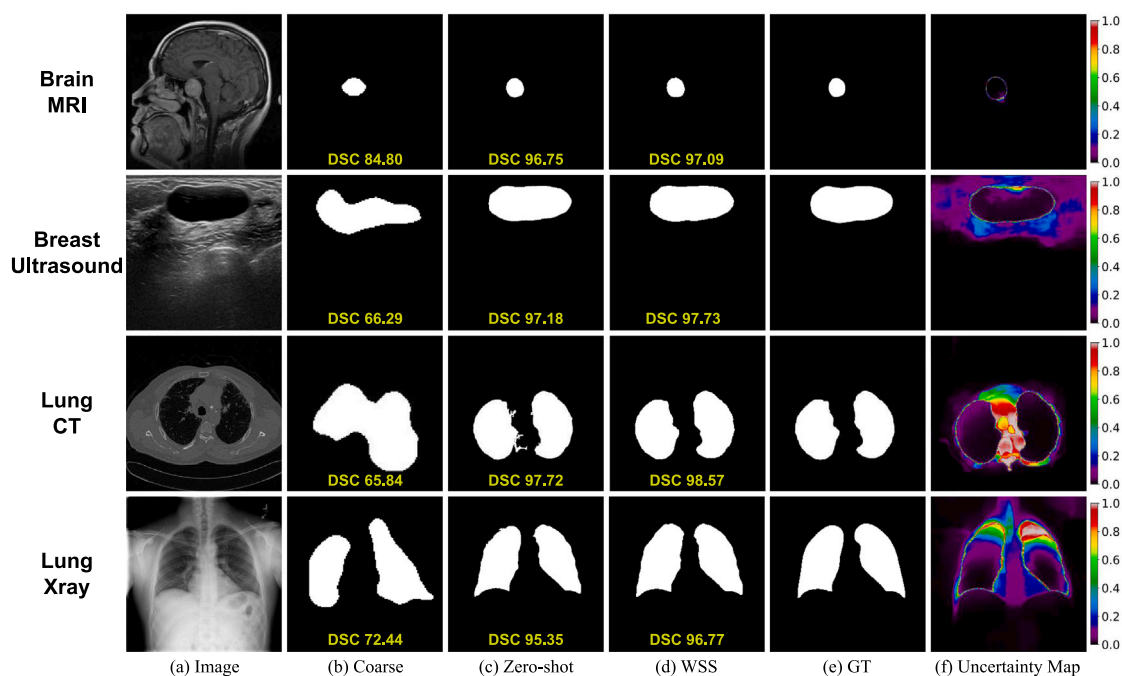


Fig. 4. Qualitative comparison of segmentation results. Coarse=post-processed saliency map, WSS=Weakly Supervised Segmentation and GT=Ground Truth. The uncertainty map corresponds to the weakly supervised segmentation.

X-ray. In Lung CT, the primary challenge for the algorithm was distinguishing between the two lobes. The post-processed results showed one large, connected contour in the center. The zero-shot refinement slightly separated these two regions, though some artifacts persisted. However, the weakly supervised training effectively corrected these false activations, producing a high-quality segmentation map. For Lung X-ray, while the weakly supervised training improved upon the less precise zero-shot masks, the improvement was not as substantial as with Lung CT. Furthermore, we also included uncertainty maps for all predictions. For Brain MRI, high uncertainty was observed only at the edges of the segmentation, which is typical. For Breast Ultrasound, high uncertainty was observed at the borders of the segmentation, while the surrounding area outside the borders showed low uncertainty. In contrast, for Lung X-rays, slight uncertainty appeared in the center of the mask, increasing towards the edges. In the case of Lung CT, high uncertainty was observed both at the edges and in the center of the lung lobes. This was largely due to the artifacts present in the zero-shot pseudo-labels.

5. Discussion

The proposed MedCLIP-SAMv2 framework demonstrates superior performance in zero-shot and weakly supervised medical image segmentation tasks than SOTA methods and the original MedCLIP-SAM method (Koleilat et al., 2024b) across four critical medical imaging modalities (CT, MRI, Ultrasound, and X-ray). By leveraging BiomedCLIP and SAM with text and visual prompts, our method exhibits robust domain and task generalization, excelling in complex tasks, such as brain and breast tumor segmentation, where smaller and intricate anatomical details pose challenges in typical segmentation tasks. Our approach notably surpasses other SOTA zero-shot and few-shot methods, especially in difficult segmentation scenarios (see Table 1). Recent methods like (Ding et al., 2022) have demonstrated the potential of CLIP for zero-shot segmentation by decoupling the pixel-level and image-level classification tasks in natural vision applications. However, such methods require fully supervised segmentation ground truths, limiting their application in settings where labels are scarce or noisy, like

medical image segmentation. In contrast, MedCLIP-SAMv2 bypasses this requirement and operates without relying on segmentation labels during training, offering a more scalable approach for medical imaging, particularly in weakly supervised settings.

Compared with the original MedCLIP-SAM, the component updates in MedCLIP-SAMv2 have greatly contributed to the performance improvement. One of the key strengths of our framework lies in the integration of M2IB for radiological tasks, which effectively extracts meaningful information from medical images and texts, enhancing segmentation performance. The introduction of the DHN-NCE loss played a crucial role in fine-tuning BiomedCLIP, enabling the model to focus on challenging details while maintaining high performance across all tasks and modalities. Importantly, the combination of M2IB and DHN-NCE allowed the model to generate coarse segmentation masks that are later refined via SAM in a zero-shot setting (see Table 5), proving the versatility of the method without the need for ground truth annotations. Finally, the effectiveness of prompt design was another critical insight. Contextually rich, descriptive prompts yielded better results in complex tasks like tumor segmentation, where finer anatomical understanding is required. Conversely, more generic prompts sufficed for simpler tasks like lung segmentation, where larger, distinct structures allowed the model to achieve strong performance with less specific guidance. This insight suggests the importance of tailoring the text prompts in visual language models for specific radiological tasks. This contrasts with findings from other studies that used the frozen BiomedCLIP encoder with an added decoder head for segmentation transfer learning, where text prompts had little impact on segmentation quality (Poudel et al., 2024). The choice of BiomedCLIP over CLIP also facilitates the success of our method. Fig. 5 shows the latent representations produced by CLIP and BiomedCLIP (both utilizing the same architecture, i.e., ViT-B/16) on sample medical images. These visualizations are generated by extracting the 196 patch tokens (excluding the CLS token) from the last layer of the image encoder, reshaping them to a $14 \times 14 \times 768$ tensor, and upsampling to the original resolution ($224 \times 224 \times 768$) using the FeatUp upsampler (Fu et al., 2024). We then apply PCA to project the 768-dimensional features into the RGB space, resulting in a $224 \times 224 \times 3$ visualization. It is important to note that these color maps are abstract and not tied to specific anatomical regions (e.g., red does not signify tumor or skull); rather, they are intended to qualitatively assess the contrast and separability in the learned representations. Compared to CLIP, BiomedCLIP exhibits greater differentiation between salient and background regions, as evidenced by the stronger visual contrast. This indicates that BiomedCLIP is more effective at capturing subtle, disease-relevant cues within medical scans using only natural language supervision, making it better suited for downstream medical imaging tasks.

Our framework's ability to operate in a weakly supervised paradigm further strengthens its potential clinical applicability. By using pseudo-labels from zero-shot segmentation to train the model from scratch, we observed notable improvements, particularly in lung CT segmentation, where the combination of zero-shot labels and weak supervision generated significant accuracy gains. To the best of our knowledge, we are the first to integrate uncertainty estimation through nnUNet with checkpoint ensembling by training on pseudo-segmentation data, providing a robust method for enhancing segmentation quality while offering insights into prediction confidence for potential end users. Uncertainty measures are essential in clinical adoption, as they help identify regions, where the model's predictions are less certain, enabling clinicians to focus on areas that may require further examination or validation.

Despite the original SAM model not being pre-trained on medical images, it showed strong performance in zero-shot settings, outperforming MedSAM and SAM-Med2D when provided with imperfect visual prompts like points and/or bounding boxes. This underscores the robustness of SAM to suboptimal input conditions as highlighted by Huang et al. (2024). Specifically, this can be seen in Fig. 4, where

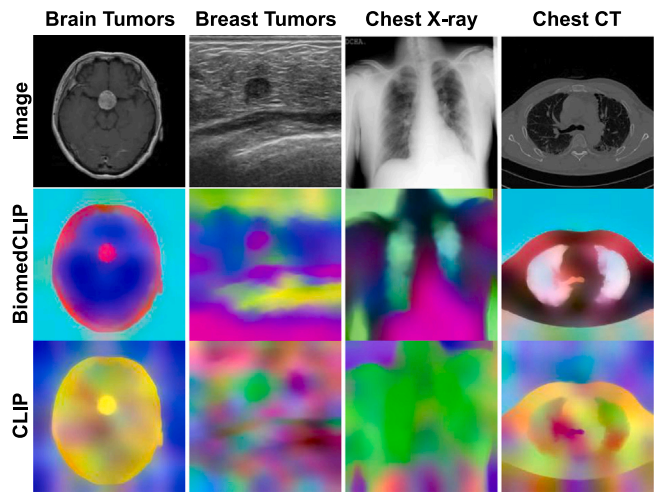


Fig. 5. Diagram showing upsampled feature representations from the last transformer layer of CLIP and BiomedCLIP. Feature Maps were upsampled using FeatUp (Fu et al., 2024) for visualization purposes.

even coarse segmentations can be refined using both zero-shot and weakly supervised methods. Although SAM demonstrates strong performance, it is important to highlight that our framework is model-agnostic and not limited to SAM. It is compatible with any interactive segmentation foundation models, provided that the visual prompts generated from CLIP saliency maps during post-processing are accurate. Notably, our framework can be seamlessly integrated with models such as ScribblePrompt (Wong et al., 2024) and SegVol (Du et al., 2024), both of which support visual prompts. This underscores the flexibility of our approach and its applicability beyond SAM-based models. Looking ahead, our future work will focus on extending our framework to handle 3D medical data, a crucial step in advancing the segmentation of volumetric imaging modalities like MRI and CT. Incorporating 3D models will enable our framework to better capture complex anatomical structures, further enhancing its clinical utility. Overall, our findings show that MedCLIP-SAMv2, with its integrated components, marks a significant step forward in the development of universal, interactive medical image segmentation. The framework's adaptability across different tasks and its ability to operate with minimal labeled data emphasize its potential for clinical adoption, particularly in resource-constrained settings. For our exploration, we focused on radiological tasks, with image modalities having more distinct characteristics than natural images. In the future, we will further incorporate and assess the performance of photograph-based biomedical images, such as histopathological images and surgical video with our proposed framework.

While our framework shows strong performance in 2D binary segmentation tasks, several limitations remain. It currently does not support multi-class segmentation, which is common in clinical applications; this could be addressed by extending the framework to handle one-vs-all class-specific prompts or by generating multi-class saliency maps with pixel-wise probabilities. Building on the insights from this study, we will explore multi-class segmentation in the future. Additionally, the framework is limited to 2D inputs due to BiomedCLIP's architecture. However, recent developments like CT-CLIP (Hamamci et al., 2024) and SAM2 (Ravi et al., 2024) offer promising opportunities for extending to 3D: for instance, a text prompt could localize an object in a single slice, while SAM2's tracking capability propagates the segmentation across adjacent slices, enabling efficient volumetric annotation. Clinically, while our method may not yet replace fully supervised systems in high-stakes diagnostic tasks, its low annotation

burden and generalizability make it suitable for interactive segmentation and rapid dataset curation, especially in the context of rare diseases where annotated data is often limited. A core motivation behind our framework is to overcome three longstanding limitations in medical deep learning: the scarcity of annotated data, limited interactivity, and poor generalizability across modalities. While most interactive segmentation systems rely on manually placed visual prompts (e.g., clicks or bounding boxes), our approach introduces a more intuitive and scalable alternative through text-based prompting. This allows users to interact with the model using natural languages, whether typed, spoken, or generated through LLMs like GPT-4, lowering the barrier for clinical adoption and reducing reliance on precise manual inputs. To explore this further, we analyzed how different prompt styles impact segmentation quality, providing insights into how clinicians might tailor prompts to specific tasks. On the interpretability aspect, we incorporated uncertainty estimation to enhance user trust and safety, which are key considerations in clinical workflows. As indicated by Zou et al. (2023), uncertainty maps help visualize prediction confidence at a local level, allowing users to identify regions where additional validation may be needed. We believe that combining intuitive interaction with transparent output interpretation offers a promising step towards more clinician-in-the-loop, trustworthy AI segmentation systems. Finally, we plan to extend our work beyond radiological tasks to other biomedical image types, such as histopathology and surgical video, to broaden the applicability of our universal, language-guided segmentation framework.

6. Conclusion

We presented MedCLIP-SAMv2, an upgraded version of the original MedCLIP-SAM framework, significantly improving segmentation performance with minimal supervision across CT, X-ray, Ultrasound, and MRI. By introducing the novel DHN-NCE loss for fine-tuning BiomedCLIP and leveraging SAM, our model achieved enhanced accuracy, particularly in complex tasks. MedCLIP-SAMv2 outperforms its predecessor through superior generalization and refined segmentation, demonstrating strong potential for clinical use in data-limited environments.

CRediT authorship contribution statement

Taha Koleilat: Writing – review & editing, Validation, Methodology, Data curation, Writing – original draft, Software, Investigation, Conceptualization. **Hojat Asgariandehkordi:** Software, Validation, Methodology. **Hassan Rivaz:** Writing – review & editing, Supervision, Methodology, Writing – original draft, Resources. **Yiming Xiao:** Writing – original draft, Resources, Writing – review & editing, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Fonds de recherche du Québec – Nature et technologies (B2X-363874).

References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
- Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A., 2020. Dataset of breast ultrasound images. Data Brief 28, 104863. <http://dx.doi.org/10.1016/j.dib.2019.104863>.
- Aleem, S., Wang, F., Maniparambil, M., Arazo, E., Dietlmeier, J., Curran, K., Connor, N.E., Little, S., 2024. Test-time adaptation with salip: A cascade of sam and clip for zero-shot medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5184–5193.
- Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., et al., 2021. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence 3 (6), e200267.
- Ayzenberg, L., Giryres, R., Greenspan, H., 2024. ProtoSAM-one shot medical image segmentation with foundational models. arXiv preprint arXiv:2407.07042.
- Bae, W., Noh, J., Kim, G., 2020. Rethinking class activation mapping for weakly supervised object localization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer, pp. 618–634.
- Baevski, A., Babu, A., Hsu, W.-N., Auli, M., 2023. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In: International Conference on Machine Learning. PMLR, pp. 1416–1429.
- Butoi, V.I., Ortiz, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2023. UniverSeg: Universal medical image segmentation. URL <https://arxiv.org/abs/2304.06131>. arXiv:2304.06131.
- Byra, M., Jarosik, P., Szubert, A., Galperin, M., Ojeda-Fournier, H., Olson, L., O’Boyle, M., Comstock, C., Andre, M., 2020. Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network. Biomed Signal Process. Control. 61.
- Chaudhry, H.A.H., Renzulli, R., Perlo, D., Santinelli, F., Tibaldi, S., Cristiano, C., Grosso, M., Limerutti, G., Fiandrotti, A., Grangotto, M., et al., 2022. Unicochest: A lung image dataset for segmentation of cancerous nodules on ct scans. In: International Conference on Image Analysis and Processing. Springer, pp. 185–196.
- Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E., 2020. Big self-supervised models are strong semi-supervised learners. Adv. Neural Inf. Process. Syst. 33, 22243–22255.
- Chen, P., Li, Q., Biaz, S., Bui, T., Nguyen, A., 2022. gScoreCAM: What is CLIP looking at? In: Proceedings of the Asian Conference on Computer Vision. ACCV.
- Chen, T., Mai, Z., Li, R., lun Chao, W., 2023. Segment anything model (SAM) enhanced pseudo labels for weakly supervised semantic segmentation. arXiv:2305.05803.
- Chen, Z., Xu, Q., Liu, X., Yuan, Y., 2024. UN-SAM: Universal prompt-free segmentation for generalized nuclei images. arXiv:2402.16663.
- Cheng, J., 2017. Brain tumor dataset. <http://dx.doi.org/10.6084/m9.figshare.1512427.v5>.
- Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K., 2023. SAM on medical images: A comprehensive study on three prompt modes. ArXiv [abs/2305.00035](https://arxiv.org/abs/2305.00035).
- Chowdhury, M.E.H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Emadi, N.A., Reaz, M.B.I., Islam, M.T., 2020. Can AI help in screening viral and COVID-19 pneumonia? IEEE Access 8, 132665–132676. <http://dx.doi.org/10.1109/ACCESS.2020.3010287>.
- Dai, H., Ma, C., Yan, Z., Liu, Z., Shi, E., Li, Y., Shu, P., Wei, X., Zhao, L., Wu, Z., Zeng, F., Zhu, D., Liu, W., Li, Q., Sun, L., Liu, S.Z.T., Li, X., 2024. SAMAug: Point prompt augmentation for segment anything model. URL <https://arxiv.org/abs/2307.01187>. arXiv:2307.01187.
- Ding, J., Xue, N., Xia, G.-S., Dai, D., 2022. Decoupling zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11583–11592.
- Du, Y., Bai, F., Huang, T., Zhao, B., 2024. Segvol: Universal and interactive volumetric medical image segmentation. Adv. Neural Inf. Process. Syst. 37, 110746–110783.
- Eslami, S., Meinel, C., de Melo, G., 2023. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In: Vlachos, A., Augenstein, I. (Eds.), Findings of the Association for Computational Linguistics: EAACL 2023. Association for Computational Linguistics, Dubrovnik, Croatia, pp. 1181–1193. <http://dx.doi.org/10.18653/v1/2023.findings-eacl.88>.
- Fu, S., Hamilton, M., Brandt, L., Feldman, A., Zhang, Z., Freeman, W.T., 2024. Featup: A model-agnostic framework for features at any resolution. arXiv preprint arXiv:2403.10516.
- Gertz, R.J., Dratsch, T., Bunck, A.C., Lennartz, S., Iuga, A.-I., Hellmich, M.G., Persigehl, T., Pennig, L., Gietzen, C.H., Fervers, P., et al., 2024. Potential of GPT-4 for detecting errors in radiology reports: Implications for reporting accuracy. Radiology 311 (1), e232714.
- Gong, S., Zhong, Y., Ma, W., Li, J., Wang, Z., Zhang, J., Heng, P.-A., Dou, Q., 2023. 3DSAM-adaptor: Holistic adaptation of SAM from 2D to 3D for promptable medical image segmentation. arXiv:2306.13465.
- Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Doga, I., Durugol, O.F., Dai, W., Xu, M., et al., 2024. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. arXiv preprint arXiv:2403.17834.

- Hu, X., Xu, X., Shi, Y., 2023. How to efficiently adapt large segmentation model(SAM) to medical images. *arXiv:2306.13731*.
- Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q., 2017. Snapshot ensembles: Train 1, get m for free. In: *The Fifth International Conference on Learning Representations*.
- Huang, Z., Liu, H., Zhang, H., Li, X., Liu, H., Xing, F., Laine, A., Angelini, E., Hendon, C., Gan, Y., 2023. Push the boundary of SAM: A pseudo-label correction framework for medical segmentation. *arXiv:2308.00883*.
- Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., Liu, S., Chi, H., Hu, X., Yue, K., Li, L., Grau, V., Fan, D.-P., Dong, F., Ni, D., 2024. Segment anything model for medical images? *Med. Image Anal.* 92, 103061. <http://dx.doi.org/10.1016/j.media.2023.103061>, URL <https://www.sciencedirect.com/science/article/pii/S1361841523003213>.
- Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* 18 (2), 203–211.
- Keicher, M., Zaripova, K., Czempiel, T., Mach, K., Khakzar, A., Navab, N., 2023. FlexR: Few-shot classification with language embeddings for structured reporting of chest X-rays. *arXiv:2203.15723*.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al., 2023. Segment anything. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4015–4026.
- Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y., 2024b. MedCLIP-SAM: Bridging text and image towards universal medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 643–653.
- Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y., 2025. Biomedcoop: Learning to prompt for biomedical vision-language models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 14766–14776.
- Konya, D., 2020. CT lung & heart & trachea segmentation.
- Krishna, S., Bhamra, N., Bleakney, R., Bhayana, R., 2024. Evaluation of reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 on a radiology board-style examination. *Radiology* 311 (2), e232715.
- Li, S., Cao, J., Ye, P., Ding, Y., Tu, C., Chen, T., 2025. Clipsam: CLIP and SAM collaboration for zero-shot anomaly segmentation. *Neurocomputing* 618, 129122.
- Li, Y., Wang, H., Duan, Y., Li, X., 2023b. CLIP surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv:2304.05653*.
- Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J., 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* 36, 28541–28564.
- Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W., 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 525–536.
- Liu, X., Huang, X., 2024. Weakly supervised salient object detection via bounding-box annotation and SAM model. *Electron. Res. Arch.* 32 (3), 1624–1645.
- Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., Lakshminarayanan, B., 2020. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Adv. Neural Inf. Process. Syst.* 33, 7498–7512.
- Loquercio, A., Segu, M., Scaramuzza, D., 2020. A general framework for uncertainty estimation in deep learning. *IEEE Robot. Autom. Lett.* 5 (2), 3153–3160.
- Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B., 2024. Segment anything in medical images. *Nature Communications* 15 (1), 654.
- Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Otsu, N., 1979. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* 9 (1), 62–66. <http://dx.doi.org/10.1109/TSMC.1979.4310076>.
- Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C., 2018. Radiology objects in context (ROCO): A multimodal image dataset. In: *CVII-STENT/LABELS@MICCAI*.
- Poudel, K., Dhakal, M., Bhandari, P., Adhikari, R., Thapaliya, S., Khanal, B., 2024. Exploring transfer learning in medical image segmentation using vision-language models. In: Burgos, N., Petitjean, C., Vakalopoulou, M., Christodoulidis, S., Coupe, P., Delingette, H., Lartizien, C., Mateus, D. (Eds.), *Proceedings of the 7nd International Conference on Medical Imaging with Deep Learning*. In: *Proceedings of Machine Learning Research*, vol. 250, PMLR, pp. 1142–1165, URL <https://proceedings.mlr.press/v250/poudel24a.html>.
- Radenovic, F., Dubey, A., Kadian, A., Mihaylov, T., Vandenhende, S., Patel, Y., Wen, Y., Ramanathan, V., Mahajan, D., 2023. Filtering, distillation, and hard negatives for vision-language pre-training. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6967–6977.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. PmlR, pp. 8748–8763.
- Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Abul Kashem, S.B., Islam, M.T., Al Maadeed, S., Zughair, S.M., Khan, M.S., Chowdhury, M.E., 2021. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Comput. Biol. Med.* 132, 104319. <http://dx.doi.org/10.1016/j.compbiomed.2021.104319>.
- Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J., 2022. Densclip: Language-guided dense prediction with context-aware prompting. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18082–18091.
- Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al., 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.
- Robinson, J., Chuang, C.-Y., Sra, S., Jegelka, S., 2021. Contrastive learning with hard negative samples. *arXiv:2010.04592*.
- Shaharabany, T., Dahan, A., Giryas, R., Wolf, L., 2023. AutoSAM: Adapting SAM to medical images by overloading the prompt encoder. *ArXiv abs/2306.06370*.
- Siragusa, I., Contino, S., La Ciura, M., Alicata, R., Pirrone, R., 2024. Medpix 2.0: A comprehensive multimodal biomedical dataset for advanced AI applications. *arXiv preprint arXiv:2407.02994*.
- Siuly, S., Zhang, Y., 2016. Medical big data: Neurological diseases diagnosis through medical data analysis. *Data Sci. Eng.* 1, 54–64.
- Taleb, A., Lippert, C., Klein, T., Nabi, M., 2021. Multimodal self-supervised learning for medical image analysis. In: *International Conference on Information Processing in Medical Imaging*. Springer, pp. 661–673.
- Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P., 2022. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* 6 (12), 1399–1406. <http://dx.doi.org/10.1038/s41551-022-00936-9>.
- Wang, Y., Rudner, T.G.J., Wilson, A.G., 2024. Visual explanations of image-text representations via multi-modal information bottleneck attribution. URL <https://arxiv.org/abs/2312.17174>. *arXiv:2312.17174*.
- Wang, Z., Wu, Z., Agarwal, D., Sun, J., 2022. MedCLIP: Contrastive learning from unpaired medical images and text. *arXiv:2210.10163*.
- Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A., 2023. Vision-language modelling for radiological imaging and reports in the low data regime. *arXiv:2303.17644*.
- Wong, H.E., Rakic, M., Guttig, J., Dalca, A.V., 2024. Scribbleprompt: Fast and flexible interactive segmentation for any biomedical image. In: *European Conference on Computer Vision*. Springer, pp. 207–229.
- Wu, Q., Zhang, Y., Elbatel, M., 2023b. Self-prompting large vision models for few-shot medical image segmentation. In: *MICCAI Workshop on Domain Adaptation and Representation Transfer*. Springer, pp. 156–167.
- Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W., 2023a. MedKLIP: Medical knowledge enhanced language-image pre-training in radiology. *arXiv:2301.02228*.
- Yang, X., Gong, X., 2024. Foundation model assisted weakly supervised semantic segmentation. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. pp. 523–532.
- Yeh, C.-H., Hong, C.-Y., Hsu, Y.-C., Liu, T.-L., Chen, Y., LeCun, Y., 2022. Decoupled contrastive learning. In: *European conference on computer vision*. Springer, pp. 668–684.
- Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Lungren, M.P., Naumann, T., Poon, H., 2023. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv:2303.00915*.
- Zhao, Y., Yang, C., Schweidtmann, A., Tao, Q., 2022. Efficient Bayesian uncertainty estimation for nnu-net. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 535–544.
- Zhou, C., Loy, C.C., Dai, B., 2022. Extract free dense labels from clip. In: *European Conference on Computer Vision*. Springer, pp. 696–712.
- Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., Fu, H., 2023. A review of uncertainty estimation and its application in medical imaging. *Meta- Radiol.* 1 (1), 100003.