Intelligent & Human Futures
WISDOM ACADEMIC https://doi.org/10.63808/ihf.v2i3.429 Volume 2 , Issue 3

Research Article

Research on Brain Tumor MRI Image Segmentation Method Based on BiomedCLIP

Mengyuan Cao[1,*], Jialu Zhao[1], Xinxin Song[1]
[1]School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 102488, China
[*]**Correspondence:** Mengyuan Cao, cmy10171007@163.com

Abstract

To address the challenges of blurry lesion boundaries, unstable training with small samples, and insufficient cross-domain generalization in brain tumor MRI image segmentation, this paper proposes BiomedCLIP-Seg, a novel segmentation method based on the biomedical vision-language pre-training model BiomedCLIP. Taking T2-FLAIR axial slices as input, the method introduces medical text semantic priors into the pixel-level lesion localization process through structured medical prompts and zero-initialized soft prompt mechanisms. Meanwhile, a cross-layer visual feature fusion module is designed to integrate shallow boundary textures, mid-level regional contexts, and high-level lesion semantics. Furthermore, a QKV cross-modal segmentation decoder is constructed to achieve explicit interaction between visual patch representations and category text semantics. To alleviate the issues of peripheral missed segmentation and region adhesion caused by weak tumor boundaries, a cross-modal consistency loss and an L2 norm-based boundary constraint loss are further introduced. Experimental results on the BraTS 2021 dataset demonstrate that BiomedCLIP-Seg achieves Dice, IoU, and HD95 scores of 0.903±0.004, 0.823±0.005, and 5.42±0.27 mm, respectively, outperforming representative methods including U-Net, TransUNet, CLIPSeg, nnU-Net, Swin UNETR, MedSAM, and SAMed. Ablation studies and hyperparameter sensitivity analyses further validate the effectiveness of structured medical prompts, soft prompt adaptation, cross-layer feature fusion, cross-modal consistency constraints, and boundary losses in improving model performance. This study provides a new technical pathway for utilizing biomedical vision-language models to tackle weak boundary segmentation problems in medical imaging.

Keywords

Medical Image Segmentation, Brain Tumor, Vision-Language Models (VLMs), Cross-modal Attention

1. Introduction

Brain tumor MRI segmentation aims to automatically localize and delineate tumors and their associated tissue regions from magnetic resonance imaging, serving as a crucial foundation for tumor volume assessment, treatment planning, and therapeutic efficacy follow-up. However, this task still faces multiple challenges: First, brain tumors exhibit significant individual variations in morphology, size, location, and signal intensity, often presenting marked intra-tumoral heterogeneity; Second, tumor boundaries are typically infiltrative and blurred, lacking clear demarcations from the surrounding normal brain parenchyma; Finally, medical image annotation is costly with limited sample sizes, and varying scanning centers and protocols further increase the difficulty of model generalization.

U-Net and its variants have been widely adopted in medical image segmentation due to their encoder-decoder architecture and skip connection mechanisms (Ronneberger et al., 2015). Subsequently, methods such as V-Net, nnU-Net, TransUNet, and Swin UNETR have continuously improved medical segmentation performance from the perspectives of 3D modeling, automated configuration, and global dependency modeling (Milletari et al., 2016; Isensee et al., 2021; Chen et al., 2021; Hatamizadeh et al., 2022). However, purely visual models primarily rely on image supervision, and their utilization of high-level medical semantics remains relatively limited. Consequently, they are prone to over-segmentation, under-segmentation, or class confusion in scenarios involving small samples and weak boundaries.

In recent years, vision-language pre-training models such as CLIP have established semantic alignment between images and texts through large-scale image-text contrastive learning, providing a new transfer learning paradigm for downstream visual tasks (Radford et al., 2021). BiomedCLIP further undergoes pre-training on large-scale biomedical image-text pairs, enabling it to encode cross-modal semantic representations that better align with medical terminology and medical image structures(Zhang et al., 2023). Compared to generic CLIP, BiomedCLIP is more suitable for medical image understanding tasks; however, its transfer mechanisms in pixel-level medical segmentation, particularly in the fine-grained segmentation of brain tumor MRI, remain to be further investigated.

To address the aforementioned issues, this paper proposes the BiomedCLIP-Seg model, specifically designed for the semantic segmentation of brain tumor MRI in the BraTS 2021 dataset. The main contributions of this work are summarized as follows:

(1) We construct a BiomedCLIP transfer framework tailored for brain tumor MRI segmentation, systematically introducing biomedical vision-language priors into

pixel-level prediction tasks for the first time to enhance the model's understanding of lesion semantics.

(2) We devise a Structured Medical Prompt and a Zero-initialized Soft Prompt mechanism, which significantly improve task adaptation efficiency and segmentation accuracy without disturbing the stability of the pre-trained backbone model.

(3) We propose a Cross-layer Visual Feature Fusion module and a QKV Cross-modal Segmentation Decoder, enabling text semantics to actively participate in lesion localization as learnable queries and achieving efficient interaction of cross-modal features.

(4) We conduct comprehensive comparative and ablation experiments on the public BraTS 2021 dataset to quantitatively validate the effectiveness of each component of the proposed method, providing new insights for the application of cross-modal priors in medical segmentation.

2. Related Work

2.1. Medical Image Segmentation Methods

Medical image segmentation methods have evolved from traditional image processing to convolutional neural networks (CNNs), and further to Transformers and foundation model adaptation. U-Net has become a classic baseline in medical segmentation tasks by employing a symmetric encoder-decoder architecture and cross-layer skip connections to integrate high-level semantics with low-level spatial details(Ronneberger et al., 2015). V-Net extends volumetric convolution operations to three-dimensional data (Milletari et al., 2016), while nnU-Net achieves stable performance across various medical segmentation tasks through automated network configuration, preprocessing, and training strategies (Isensee et al., 2021).

The Transformer architecture models long-range dependencies through self-attention mechanisms, providing global context modeling capabilities for medical image segmentation. TransUNet integrates a Transformer encoder with a U-Net-style decoder (Chen et al., 2021), while Swin UNETR further employs a hierarchical window attention structure to model three-dimensional MRI volumetric data (Hatamizadeh et al., 2022). Additionally, PSPNet aggregates multi-scale contextual information via a pyramid pooling module (Zhao et al., 2017). Although these methods enhance the modeling capacity for complex structures, they typically still rely on visual supervision and lack sufficient utilization of medical text priors.

Recently, data-efficient learning and segmentation foundation models have emerged as pivotal research directions in medical image segmentation. SSCLMix enhances representation robustness through self-supervised contrastive learning combined with mixed data

augmentation (Hou et al., 2025). The Segment Anything Model (SAM) has demonstrated the formidable capabilities of promptable segmentation foundation models (Kirillov et al., 2023). Furthermore, MedSAM and SAMed have advanced the application of foundation models in medical segmentation from the perspectives of medical image adaptation and parameter-efficient fine-tuning, respectively (Ma et al., 2024; Zhang & Liu, 2023). Therefore, this paper incorporates these representative methods into our comparative experiments to broaden the experimental coverage and enhance the credibility of the conclusions.

2.2. Vision-Language Pre-training and Prompt Learning

Vision-language pre-training establishes a cross-modal semantic space through image-text pairs, enabling models to utilize natural language descriptions to guide visual recognition. The success of CLIP demonstrates that text can serve not only as an extension of class labels but also as a semantic interface for downstream tasks (Radford et al., 2021). Building upon this, CLIPSeg further introduces text prompts into image segmentation, transferring the image-text alignment capabilities to dense prediction tasks via a lightweight decoder (Zhang et al., 2022).

Research on prompt learning has demonstrated that the construction of prompt templates significantly impacts the transfer effectiveness of vision-language models. CoOp extends discrete prompts into learnable context vectors (Zhou et al., 2022), while Visual Prompt Tuning (VPT) achieves parameter-efficient adaptation through a small number of trainable visual prompt tokens (Jia et al., 2022). In the brain tumor segmentation scenario, this paper introduces both structured medical prompts and zero-initialized soft prompts to balance medical semantic expression with training stability under limited annotation conditions.

2.3. Biomedical Vision-Language Models

BiomedCLIP is a vision-language foundation model tailored for the biomedical domain, pre-trained via contrastive learning on large-scale biomedical image-text pairs covering diverse medical imaging modalities, including radiology scans, microscopic images, and pathological slides (Zhang et al., 2023). Compared to general-purpose vision-language models, both its text encoder and visual encoder are better aligned with medical terminology, anatomical structures in medical images, and biomedical semantic representations, thereby making it more suitable for transfer to downstream medical image analysis tasks.

Existing research related to BiomedCLIP has predominantly focused on tasks such as medical image classification, image-text retrieval, and visual question answering. However, translating its inherent biomedical semantic priors into pixel-level segmentation capabilities

requires further exploration. To address this, this paper introduces structured medical prompts, zero-initialized soft prompts, cross-layer visual feature fusion, and a cross-modal decoding mechanism. These strategies enable the image-text alignment capabilities of BiomedCLIP to more effectively serve the task of weak-boundary brain tumor MRI segmentation.

3. Methodology

3.1. Overall Framework

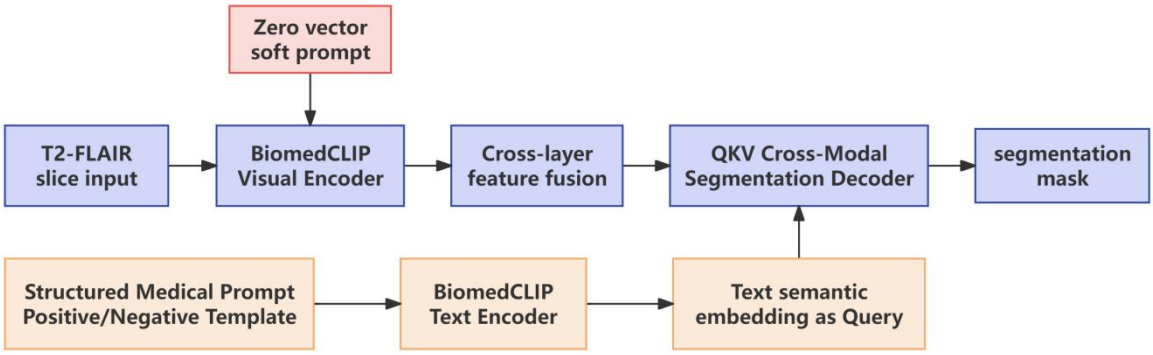
To address the limitations of traditional medical image segmentation methods, which primarily rely on visual features and lack medical semantic guidance—often resulting in blurred lesion boundaries, confusion between normal tissues and tumor regions, and unstable training with small samples—we propose a brain tumor MRI image segmentation framework based on BiomedCLIP, termed BiomedCLIP-Seg. This framework takes T2-FLAIR axial slices as input, normalizes the single-channel MRI images, duplicates them into three channels, and feeds them into the BiomedCLIP visual encoder to extract visual patch representations. Concurrently, structured medical text descriptions are constructed for the segmentation categories, which are then processed by the BiomedCLIP text encoder to obtain class-level medical semantic embeddings.

In the visual branch, we insert zero-initialized learnable tokens between the class token and patch tokens to achieve parameter-efficient adaptation while maintaining the stability of the pre-trained vision-language space. Subsequently, patch tokens are extracted from multiple Transformer blocks and integrated via a cross-layer visual feature fusion module to combine low-level boundary textures, mid-level regional context, and high-level semantic information. Finally, we employ a QKV cross-modal segmentation decoder to facilitate interaction between the fused visual features and the class text semantic embeddings. This generates semantically enhanced spatial response maps, which are then processed by a lightweight decoder to output the brain tumor segmentation masks.

Overall, BiomedCLIP-Seg comprises four core components: the construction of structured medical text descriptions, visual encoding with zero-initialized learnable token adaptation, cross-layer visual feature fusion, and QKV cross-modal segmentation decoding. This design enables the transfer of class-level semantic knowledge from biomedical vision-language pre-trained models to pixel-level segmentation tasks, thereby enhancing the localization capability for weak boundary regions in brain tumors. The overall architecture is illustrated in Figure 1.

Figure 1

The overall framework of BiomedCLIP-Seg



Note. Overall architecture of the BiomedCLIP-driven cross-modal medical image segmentation framework. T2-FLAIR brain slices are input to the BiomedCLIP visual encoder augmented with a zero vector soft prompt for multi-scale visual feature extraction, followed by cross-layer feature fusion. Structured positive and negative medical prompt templates are processed by the BiomedCLIP text encoder to produce text semantic embeddings that act as queries. The QKV cross-modal segmentation decoder integrates visual features and text query embeddings to output the final segmentation mask.

3.2. Construction of Structured Medical Prompts

To fully leverage the textual semantic capabilities of BiomedCLIP, this paper avoi ds using simple single category labels as prompts. Instead, we construct a structured medical prompt set for each segmentation category. Let the segmentation task consist of categories, including the background and several tumor sub-regions, such as the bac kground area, peritumoral edema, tumor core, and enhancing tumor. For the category, we construct a corresponding medical prompt set to describe the radiological characteri stics of that category in brain MRI images.

For instance, for the peritumoral edema region, we construct prompts such as “ax ial T2-FLAIR brain MRI with peritumoral edema region” and “hyperintense edema are a surrounding brain tumor”; for the tumor core region, prompts include “brain MRI w ith tumor core region” and “abnormal glioma core tissue in axial FLAIR MRI”; and f or the background or normal brain tissue, prompts are constructed as “normal brain pa renchyma on axial FLAIR MRI” and “background or non-tumor tissue in brain MRI”. By constructing medical semantic prompts tailored to different categories, we enable t he text side to explicitly encode the categorical distinctions among different tissue regi ons.

Let denote the BiomedCLIP text encoder. The text semantic embedding for the ca tegory is defined as:

$$e_c=\frac{1}{|T_c|}\sum_{i=1}^{|T_c|}E_t\left(t_{c,i}\right),\quad c=1,2,\ldots,C \tag{1}$$

The text embeddings for all categories can be represented as:

$$E=\left[e_1;e_2;\cdots;e_C\right]\in\mathbb{R}^{C\times d} \tag{2}$$

By averaging the embeddings of multiple prompt templates within the same category, we reduce the instability caused by the wording of a single prompt. This allows the text side to simultaneously encode rich medical contextual information—such as "normal brain tissue", "edema", "tumor core" and "enhancing regions"—thereby providing robust semantic constraints for subsequent multi-class pixel localization.

3.3. Visual Encoding and Zero-initialized Soft Prompts

Let the input 2D T2-FLAIR slice be . After grayscale normalization and three-channel replication, we obtain:

$$X'\in\mathbb{R}^{3\times H\times W} \tag{3}$$

In this study, we adopt BiomedCLIP-PubMedBERT_256-vit_base_patch16_224 as the vision-language backbone model. Specifically, the visual encoder is a ViT-B/16 comprising 12 Transformer blocks with 12 attention heads and a hidden dimension of 768, utilizing a patch size of 16×16. Given an input resolution of 224×224, the visual encoder generates patch tokens. The text encoder is based on PubMedBERT, which supports a maximum sequence length of 256 and outputs embeddings with a dimensionality of 768.

To achieve stable adaptation under limited medical annotations, we insert trainable tokens between the class token and the patch tokens. Letting denote the feature dimension of the tokens, the learnable token matrix is defined as. In this study, we set and , and employ zero initialization:

$$Z_0=[x_{cls};P;x_1;x_2;\ldots;x_N],\quad P_0=\mathbf{0}\in\mathbb{R}^{8\times 768} \tag{4}$$

where denotes the class token and represents the patch token. Zero initialization ensures that the model maintains consistency with the original BiomedCLIP representation space during the early stages of training, thereby avoiding representation perturbations caused by random initialization. During the training process, we freeze the first 8 Transformer blocks of the visual encoder and only update the learnable tokens, LayerNorm layers, the last 4 Transformer blocks, the cross-layer fusion module, and the segmentation decoder. This strategy is adopted to mitigate the risk of overfitting and improve training stability in few-shot scenarios.

3.4. Cross-layer Feature Fusion and QKV Cross-Modal Decoder

Intelligent & Human Futures
WISDOM ACADEMIC https://doi.org/10.63808/ihf.v2i3.429 Volume 2 , Issue 3

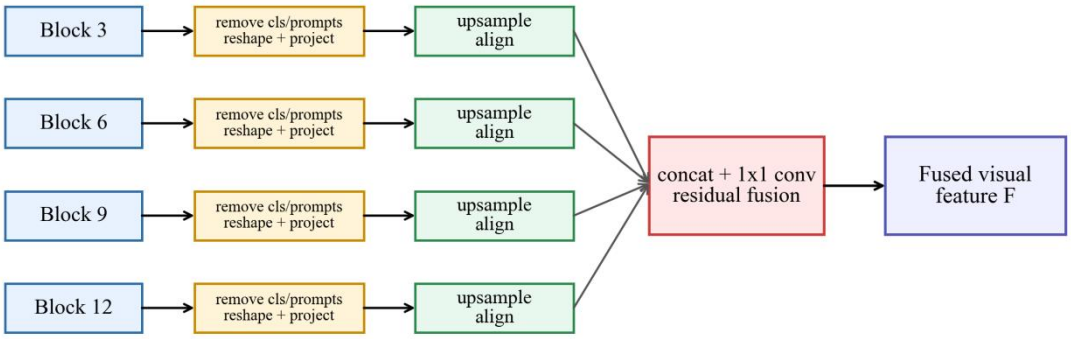
As shown in Figure 2, the tokens from the final layer of a Vision Transformer (ViT) often possess strong semantic abstraction capabilities but lack sufficient spatial details, which is detrimental to recovering fine boundaries in multi-class brain tumor segmentation. To enhance the model's utilization of visual information at different levels, this paper extracts patch tokens from the 3rd, 6th, 9th, and 12th Transformer blocks. After removing the class token and soft prompt tokens, they are reshaped into 2D feature maps, and multi-scale features are obtained through linear projection and upsampling. The cross-layer fusion process can be formulated as:

$$F_l=Reshape(Proj(Z_l^{patch})),l\in\{3,6,9,12\} \quad (5)$$

where denotes the patch tokens from the output of the Transformer block, and is the linear projection operation. Subsequently, features from different levels are aligned in channels, upsampled, and fused to obtain the fused visual feature $F\in\mathbb{R}^{N\times d}$.

Figure 2

Structure of the Cross-layer Visual Feature Fusion module



Note. Architecture of the cross-layer feature fusion module. Feature outputs from Block 3, Block 6, Block 9, and Block 12 of the BiomedCLIP visual encoder are processed by removing cls and prompt tokens, reshaping, and linear projection. Each feature branch undergoes upsampling and spatial alignment. The aligned multi-scale features are concatenated and processed via a 1x1 convolution with residual fusion to generate the final fused visual feature F.

In the cross-modal decoding stage, we use the multi-class text embeddings as the Query, and the fused visual features as the Key and Value, thereby establishing a correspondence between category semantics and image spatial regions:

$$Q=EW_Q,\quad K=FW_K,\quad V=FW_V \quad (6)$$

$$A=Softmax\left(\frac{QK^T}{\sqrt{d}}\right),O=AV \quad (7)$$

where represents the correlation weights between each visual patch and the medical semantics of various categories, and denotes the cross-modal visual features at the patch level

enhanced by semantic information. Since O retains the patch dimension, it can be reshaped into a spatial feature map to predict a coarse-grained segmentation response:

$$M_0=Softmax(Conv(Reshape(O))) \tag{8}$$

Finally, the decoder progressively fuses with the cross-layer visual features and upsamples them to the original input resolution to obtain the predicted mask hat $\hat{Y}$. This design enables medical text semantics to participate in pixel-level localization as category priors, which helps reduce false positives caused by high-signal regions in normal brain tissues and improves segmentation continuity at weak tumor boundaries.

3.5. Loss Function

To simultaneously optimize region overlap, pixel-wise classification, cross-modal semantic consistency, and boundary integrity, this paper adopts a composite loss function. Let the ground-truth multi-class segmentation label be and the prediction result be . The primary segmentation loss consists of the multi-class Dice loss and the multi-class Cross-Entropy loss:

$$\mathcal{L}_{seg}=\mathcal{L}_{Dice}(\hat{Y},Y)+\mathcal{L}_{CE}(\hat{Y},Y) \tag{9}$$

where the multi-class Dice loss can be formulated as:

$$Dice=1-\frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_i \hat{Y}_{c,i}Y_{c,i}+\epsilon}{\sum_i \hat{Y}_{c,i}+\sum_i Y_{c,i}+\epsilon} \tag{10}$$

To enhance the consistency between visual features and category-level text semantics, we further introduce a category-level cross-modal consistency loss. Let denote the visual feature pooled from the predicted region of the c -th category, and be the corresponding category text embedding. To ensure that the visual feature of the c category is closer to its corresponding text semantics while being distant from other category text semantics, the cross-modal consistency loss is defined as:

$$\mathcal{L}_{cm}=\frac{1}{C}\sum_{c=1}^{C}\left[max(0,\delta-cos(v_c,e_c))+\max_{k\neq c}cos(v_c,e_k)\right] \tag{11}$$

where is the margin coefficient, and $\cos(\cdot,\cdot)$ denotes the cosine similarity. This loss constrains the visual representations of different category regions to remain consistent with their corresponding medical semantics, thereby reducing category confusion in multi-class segmentation.

Furthermore, to alleviate the issues of outer edge omission and boundary adhesion between classes caused by blurred brain tumor boundaries, this paper introduces a boundary constraint loss L_{edge}. This term imposes constraints by comparing the differences in boundary gradients between the predicted mask and the ground-truth mask, which can be expressed as:

$$L_{\text{edge}}=\frac{1}{C}\sum_{c=1}^{C}\left\|\nabla\hat{Y}_c-\nabla Y_c\right\|_2^2 \tag{12}$$

The final optimization objective is formulated as:

$$L=L_{seg}+\lambda_{cm}L_{cm}+\lambda_{edge}L_{edge} \quad (13)$$

In this paper, we set λ_{cm}=0.1、λ_{edge}=0.05, and verify the stability of these weights on the validation set. By jointly optimizing the multi-class segmentation loss, the category-level cross-modal consistency loss, and the boundary constraint loss, the model is able to improve the discrimination capability among different tumor sub-regions and the boundary recovery accuracy, while maintaining overall regional segmentation performance.

4. Experiments and Results Analysis

4.1. Datasets and Preprocessing

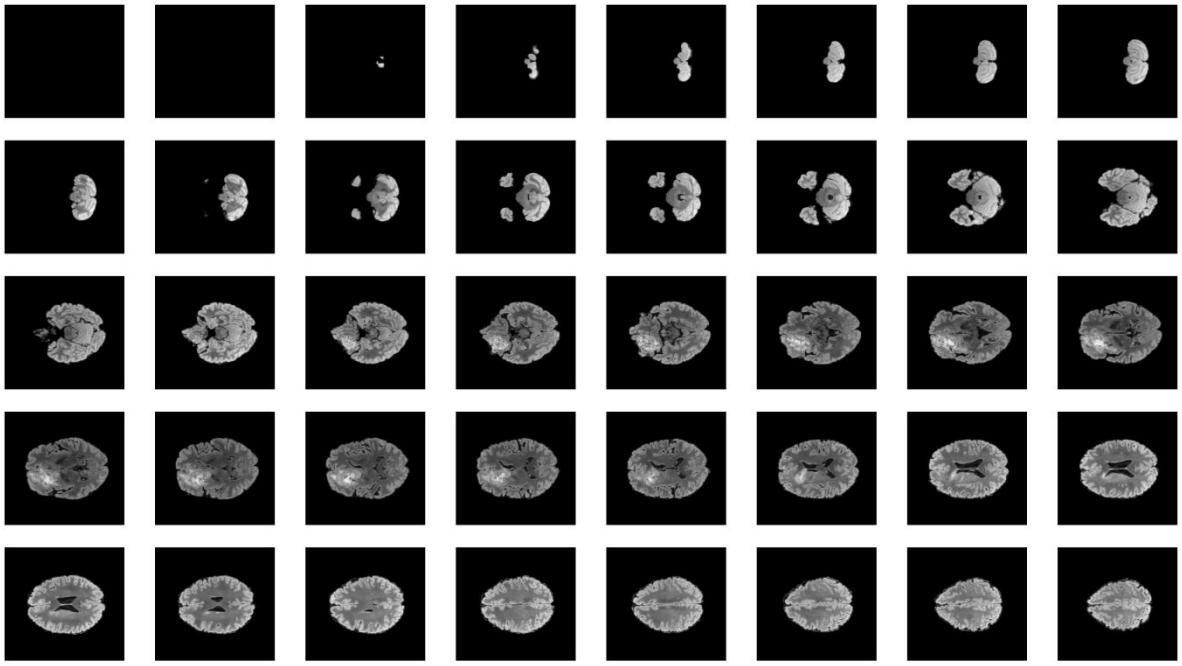
This study employs the BraTS 2021 brain tumor MRI dataset to validate the generalization capability and robustness of the proposed model in brain tumor segmentation tasks. The dataset comprises glioma MRI scans collected from multiple institutions, using diverse equipment and scanning protocols. It provides four modalities—T1, T1Gd, T2, and T2-FLAIR—along with corresponding expert annotations. In this work, we utilize only the T2-FLAIR sequence for the 2D segmentation task, with label categories including the background, peritumoral edema, tumor core, and enhancing tumor.

Since the BiomedCLIP visual encoder is inherently pre-trained as a 2D Vision Transformer (ViT), directly extending it to 3D volumetric data would require substantial architectural redesign and incur significantly higher GPU memory overhead. Therefore, this study adopts a 2D axial slice-level setting. It should be noted that this simplification may compromise the modeling of inter-slice continuity along the Z-axis and affect three-dimensional consistency in clinical volumetric assessments. Consequently, the results presented herein should be interpreted as 2D slice-level segmentation performance, and future work will further explore multimodal 3D architectures.

Figure 3

The BraTS 2021 brain tumor MRI dataset

Intelligent & Human Futures
WISDOM ACADEMIC https://doi.org/10.63808/ihf.v2i3.429 Volume 2 , Issue 3



Note. Sequential axial slices of T2-FLAIR brain MRI scans. The subfigures display continuous cross-sectional images sampled from inferior to superior brain regions, illustrating the gradual anatomical variation of brain tissue along the axial direction.

To ensure consistency in experimental settings and comparability of results, all raw images undergo a standardized preprocessing pipeline. First, z-score normalization is applied to the non-zero brain region voxels of each case to eliminate intensity discrepancies across different scans. Second, the 3D volumetric data is unfolded into 2D axial slices. Third, slices containing tumors are retained, and a certain proportion of tumor-free slices are randomly sampled during the training phase to balance foreground and background classes. Finally, both images and labels are resized to a resolution of 224×224 pixels. The dataset is divided at the patient level into training, validation, and testing sets consisting of 1000, 125, and 126 cases, respectively, to prevent data leakage caused by random slice-level partitioning. This rigorous data construction and preprocessing pipeline lays a solid foundation for the subsequent quantitative evaluation and qualitative analysis. A schematic illustration of the dataset is shown in Figure 3.

4.2. Experimental Settings

All methods were evaluated under the same training, validation, and test splits. Training was performed using the AdamW optimizer (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) with a batch size of 16 for a total of 120 epochs. A linear warmup strategy was applied during the first 10 epochs, followed by a cosine decay learning rate schedule. The learning rate for the last four Transformer blocks in the visual encoder was set to 0.00001, while the soft

prompts, cross-layer feature fusion module, and segmentation decoder used a learning rate of 0.0001. The weight decay was configured to 0.0001. Data augmentation techniques included random flipping, random rotation, scaling, gamma perturbation, and mild elastic deformation. All experiments were repeated across three random seeds, and both the mean and standard deviation are reported. Unless otherwise specified, the number of soft prompt tokens was fixed at m = 8.

The evaluation metrics comprise the Dice coefficient, mean Intersection over Union (IoU), and 95th percentile Hausdorff Distance (HD95). Specifically, Dice and IoU are employed to assess regional overlap, where higher values indicate superior performance. Conversely, HD95 quantifies the distance deviation between the predicted and ground-truth boundaries in millimeters; lower values correspond to more precise boundary localization. Incorporating the HD95 metric facilitates a more direct validation of the proposed method's effectiveness in addressing the weak-boundary segmentation challenge inherent in brain tumors.

4.3. Comparative Experiments

Table 1 presents the quantitative performance comparison between BiomedCLIP-Seg and representative baseline methods on the BraTS 2021 dataset. Beyond U-Net, TransUNet, and CLIPSeg, we further include nnU-Net, Swin UNETR, MedSAM, and SAMed to provide comprehensive coverage of classical convolutional networks, Transformer-based segmentation architectures, automated medical segmentation frameworks, and recent adaptation approaches for medical segmentation foundation models.

Table 1

Comparison of segmentation results among different methods (mean ± std)

Method	Dice	IoU	HD95（mm）↓
U-Net	0.871 ± 0.008	0.773 ± 0.010	8.72 ± 0.41
TransUNet	0.887 ± 0.006	0.797 ± 0.008	7.11 ± 0.36
CLIPSeg	0.833 ± 0.010	0.715 ± 0.012	10.94 ± 0.55
nnU-Net	0.891 ± 0.006	0.803 ± 0.007	6.67 ± 0.34
Swin UNETR	0.895 ± 0.005	0.811 ± 0.006	6.24 ± 0.29
MedSAM	0.884 ± 0.007	0.792 ± 0.008	6.95 ± 0.38
SAMed	0.887 ± 0.005	0.805 ± 0.006	6.63 ± 0.31
BiomedCLIP-Seg	0.903 ± 0.004	0.823 ± 0.005	5.42 ± 0.27

Note. Lower values of HD95 indicate better segmentation performance. w/o = without. Dice = Dice similarity coefficient; IoU = intersection over union; HD95 = 95% Hausdorff distance.

As shown in Table 1, BiomedCLIP-Seg achieves the best performance across all three evaluation metrics, namely Dice coefficient, IoU, and HD95, with scores of 0.903±0.004, 0.823±0.005, and 5.42±0.27 mm, respectively. Compared to U-Net, the proposed method yields a 0.032 improvement in Dice score and a 3.30 mm reduction in HD95, demonstrating that vision-language pre-training and cross-modal prompting can provide effective priors for brain tumor localization and boundary recovery. Furthermore, compared to the strongest recent baseline SAMed, our method still achieves a 0.016 increase in Dice and a 1.21mm decrease in HD95, indicating that biomedical domain-specific pre-training combined with cross-layer visual feature fusion is better suited for identifying abnormal regions in MRI scans.

The performance of CLIPSeg falls below that of most medical segmentation baselines, indicating significant limitations in directly transferring general text-guided segmentation models to specialized medical scenarios such as brain tumor MRI. In contrast, BiomedCLIP-Seg leverages biomedical vision-language pre-training knowledge, structured medical prompts, and a boundary-constrained loss function, achieving more stable performance in both regional overlap and boundary localization.

4.4. Hyperparameter Sensitivity Analysis

To evaluate the impact of the two loss balancing coefficients and the soft prompt length on model performance, we varied λ_{cm} , and m while keeping all other training settings unchanged. The results are presented in Table 2.

Table 2

Hyperparameter sensitivity analysis on the validation set

Setting	Dice	IoU	HD95（mm）↓
λ_{cm}=0.05，λ_{edge}=0.05，m=8	0.899	0.816	5.78
λ_{cm}=0.10，λ_{edge}=0.05，m=8	0.903	0.823	5.42
λ_{cm}=0.20，λ_{edge}=0.05，m=8	0.901	0.820	5.51
λ_{cm}=0.10，λ_{edge}=0.025，m=8	0.900	0.818	5.69
λ_{cm}=0.10，λ_{edge}=0.10，m=8	0.902	0.821	5.48
λ_{cm}=0.10，λ_{edge}=0.05，m=4	0.899	0.817	5.74
λ_{cm}=0.10，λ_{edge}=0.05，m=16	0.901	0.820	5.56

Note. Lower values of HD95 indicate better segmentation performance. Dice = Dice similarity coefficient; IoU = intersection over union; HD95 = 95% Hausdorff distance. λ_{cm}= weight of cross-modal consistency loss; λ_{edge} = weight of boundary-constrained loss; m = number of prompt tokens.

As can be seen from Table 2, the model exhibits good stability around the default configuration. It achieves the best overall performance in terms of Dice score, IoU, and HD95 when λ_{cm}=0.10，λ_{edge}=0.05， m=8. If is too small, the consistency constraint between visual regional features and medical text semantics becomes insufficient; conversely, if it is too large, it may overemphasize semantic alignment at the expense of pixel-level segmentation details. Similarly, an appropriate boundary loss helps reduce HD95, whereas an excessively large may cause the model to overly focus on edge gradients, thereby compromising the prediction stability within the target regions.

Regarding the soft prompt length, when m=4, the capacity of learnable contextual tokens is insufficient. Conversely, while increasing m to 16 raises the number of parameters, it yields limited performance improvements and may even introduce redundant prompt tokens. Therefore, we set m=8 as a trade-off between performance and parameter efficiency.

4.5. Ablation Study

To verify the contribution of each module to model performance, we conducted ablation experiments under identical experimental settings. The results are presented in Table 3.

Table 3

Results of the ablation study

Variant	Dice	IoU	HD95（mm）↓
Full Model	0.903	0.823	5.42
w/o Structured Prompt (Single Template Only)	0.894	0.808	5.91
w/o Cross-layer Feature Fusion (Last-layer Tokens Only)	0.897	0.812	5.76
Frozen Visual Backbone (Decoder and Prompts Only)	0.891	0.804	6.08
w/o Cross-modal Consistency Loss	0.899	0.817	5.62
w/o Boundary-constrained Loss	0.900	0.819	5.86

Note. Lower values of HD95 indicate better segmentation performance. w/o = without. Dice = Dice similarity coefficient; IoU = intersection over union; HD95 = 95% Hausdorff distance.

As shown in Table 3, freezing the entire visual backbone leads to the most significant performance degradation, indicating that brain tumor segmentation tasks still require limited

fine-tuning of high-level visual representations. Removing the structured prompt reduces the Dice score from 0.903 to 0.894, demonstrating that medical semantic descriptions can enhance the model's ability to distinguish between tumor regions and normal brain tissue. Furthermore, eliminating cross-layer feature fusion increases the HD95 from 5.42 mm to 5.76 mm, suggesting that the integration of multi-level visual features plays a crucial role in recovering weak boundaries.

Furthermore, removing either the cross-modal consistency loss or the boundary-constrained loss leads to a performance decline. Notably, eliminating the boundary-constrained loss results in a significant increase in HD95, indicating that the L2-norm-based boundary gradient constraint can effectively reduce the discrepancy between the predicted and ground-truth boundaries.

4.6. Visualization Analysis

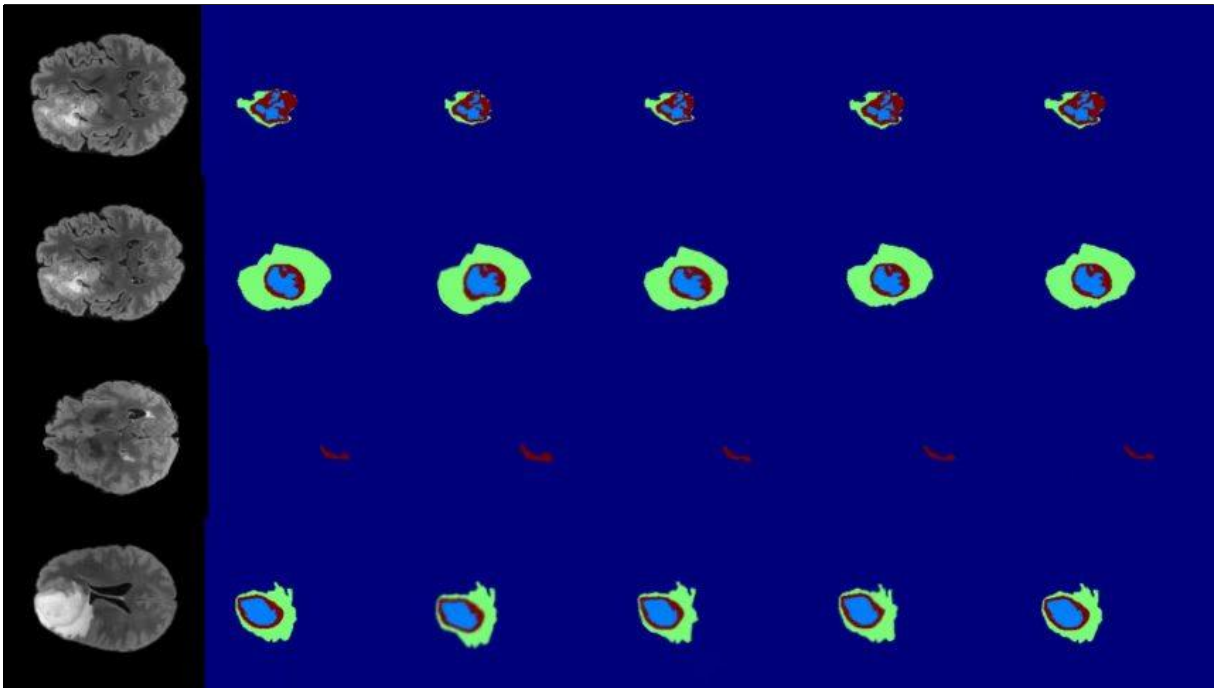
As shown in Figure 4, four representative brain tumor MRI segmentation cases are selected for visual comparison, including the original images, ground-truth annotations, and the prediction results of nnU-Net, MedSAM, Swin UNETR, and BiomedCLIP-Seg. Overall, BiomedCLIP-Seg shows better visual consistency in terms of tumor shape recovery, boundary alignment, and hierarchical identification of different subregions, and its prediction results are the closest to the ground truth.

Specifically, for samples with small tumor regions and weak boundaries, some comparison methods exhibit varying degrees of missed segmentation. This is particularly evident in small lesions or regions with unclear edges, where incomplete target regions, broken contours, or local omissions may occur. In contrast, BiomedCLIP-Seg better preserves the overall tumor structure and maintains a more continuous segmentation contour in weak-boundary areas, indicating its stronger ability to capture fine-grained boundary information.

Figure

Visual comparison (From left to right: Original image, Ground truth, nnU-Net prediction, MedSAM prediction, Swin UNETR prediction and BiomedCLIP-Seg prediction.)

Intelligent & Human Futures
WISDOM ACADEMIC https://doi.org/10.63808/ihf.v2i3.429 Volume 2 , Issue 3



Note. Qualitative segmentation visualization for brain tumor subregions. The left column displays original T2-FLAIR MRI slices. The columns on the right show segmentation results from different experimental variants; blue represents the enhancing tumor region, red indicates the necrotic tumor core, and green denotes peritumoral edema.

5. Conclusion

In this paper, we propose BiomedCLIP-Seg, a brain tumor MRI image segmentation method based on BiomedCLIP. This method leverages structured medical prompts and zero-initialized soft prompts to enhance the adaptability of BiomedCLIP for brain tumor segmentation tasks. Furthermore, it translates medical text semantics into pixel-level lesion localization capabilities through cross-layer visual feature fusion and a QKV cross-modal segmentation decoder. Meanwhile, a cross-modal consistency loss and an L2-norm-based boundary-constrained loss are introduced to improve regional semantic consistency and weak boundary localization accuracy.

Experimental results on the BraTS 2021 dataset demonstrate that BiomedCLIP-Seg outperforms representative methods, including U-Net, TransUNet, CLIPSeg, nnU-Net, Swin UNETR, MedSAM, and SAMed, in terms of Dice, IoU, and HD95 metrics. Furthermore, ablation studies confirm that structured prompts, cross-layer visual feature fusion, soft prompt tuning, cross-modal consistency constraints, and boundary-constrained loss all contribute positively to the overall model performance.

It should be noted that the current implementation still relies on 2D axial slices as input, which fails to fully exploit the inter-slice continuity within 3D MRI volumetric data. This

limitation may lead to discontinuous predictions along the Z-axis and compromise the accuracy of clinical volume assessments. Future research will further extend this work to multi-modal 3D segmentation, cross-center generalization evaluation, and more efficient clinical deployment strategies. Additionally, we plan to systematically evaluate the impact of 3D modeling on clinical volume measurements and diagnostic decision-making.

Acknowledgments

The authors have no additional acknowledgments to declare.

Author contributions

Conceptualization, Mengyuan Cao and Jialu Zhao; methodology, Mengyuan Cao; software, Mengyuan Cao and Jialu Zhao; validation, Mengyuan Cao, Jialu Zhao and Xinxin Song; formal analysis, Mengyuan Cao; investigation, Mengyuan Cao and Jialu Zhao; resources, Mengyuan Cao and Xinxin Song; data curation, Mengyuan Cao and Jialu Zhao; writing—original draft preparation, Mengyuan Cao; writing—review and editing, Mengyuan Cao, Jialu Zhao and Xinxin Song; visualization, Mengyuan Cao; supervision, Mengyuan Cao; project administration, Mengyuan Cao; funding acquisition, Mengyuan Cao. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the 2026 Undergraduate Innovation and Entrepreneurship Training Program of Beijing Technology and Business University, grant number S202610011006.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Conflict of interest

The authors declare that there is no conflict of interest.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

All authors have given their consent.

Additional information

Received: 25 May 2026

Accepted: 18 June 2026

Published online: 16 August 2026

Publisher's Note

References

[1] Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F. C., Pati, S., Prevedello, L. M., Rudie, J. D., Sako, C., Shinohara, R. T., Bergquist, T., Chai, R., Eddy, J., … Bakas, S. (2021). The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2107.02314

[2] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J. S., Freymann, J. B., Farahani, K., & Davatzikos, C. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data, 4*, Article 170117. https://doi.org/10.1038/sdata.2017.117

[3] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2102.04306

[4] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*(3), 297–302. https://doi.org/10.2307/1931617

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *Proceedings of the 9th International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2010.11929

[6] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., & Xu, D. (2022). Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Vol. 12962, pp. 272–284). Springer. https://doi.org/10.1007/978-3-031-08999-2_22

[7] Hou, J., Zhou, Y., Qin, J., Zhang, J., Li, X., & Gao, S. (2025). SSCLMix: A self-supervised contrastive learning-based data mixing augmentation method. *Neural Networks, 192*, Article 108171. https://doi.org/10.1016/j.neunet.2025.108171

[8] Isensee, F., Jaeger, P. F., Kohl, S. A. 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. https://doi.org/10.1038/s41592-020-01008-z

[9] Jia, M., Tang, L., Chen, B. C., Cardie, C., & Belongie, S. (2022). Visual prompt tuning. In *Proceedings of the European Conference on Computer Vision* (Vol. 13693, pp. 709–727). Springer. https://doi.org/10.1007/978-3-031-19827-4_41

[10] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1412.6980

[11] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 4015–4026). IEEE. https://doi.org/10.1109/ICCV51070.2023.00344

[12] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1711.05101

[13] Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). Segment anything in medical images. *Nature Communications, 15*(1), Article 654. https://doi.org/10.1038/s41467-024-44940-2

[14] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., … Van Leemput, K. (2015). The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging, 34*(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

[15] Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)* (pp. 565–571). IEEE. https://doi.org/10.1109/3DV.2016.79

[16] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748–8763). PMLR. https://doi.org/10.48550/arXiv.2103.00020

[17] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention* (Vol. 9351, pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[18] Zhang, K., & Liu, D. (2023). Customized segment anything model for medical image segmentation [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2304.13785

[19] Lüddecke, T., & Ecker, A. (2022). CLIPSeg: Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7086–7096). IEEE. https://doi.org/10.1109/CVPR52688.2022.00695

[20] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., … Poon, H. (2023). BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.00915

[21] Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision, 130*, 2337–2348. https://doi.org/10.1007/s11263-022-01653-1

[22] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2881–2890). IEEE. https://doi.org/10.1109/CVPR.2017.660