BIRES: BIDIRECTIONAL SEMANTIC RECALIBRATION FOR TEXT-GUIDED MULTI-SCALE MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image segmentation remains challenging due to substantial variations in target scale, complex and poorly defined boundaries, and the limited semantic awareness of vision-only models. Although multimodal approaches introduce complementary semantic priors, the granularity gap between global textual semantics and dense visual representations complicates effective cross-modal fusion. Meanwhile, many existing methods treat language primarily as auxiliary guidance for visual features, with limited image-conditioned feedback to textual representations, resulting in asymmetric cross-modal interaction. To address these limitations, we propose **BiReS**, a multimodal segmentation framework that combines bidirectional semantic recalibration with scale-aware feature refinement. Specifically, we introduce a **Semantic Recalibrator (SR)** that leverages global textual semantics to modulate visual responses along both spatial and channel dimensions, while attention-weighted visual pooling reciprocally updates textual representations, forming a closed-loop semantic recalibration process. To handle target scale heterogeneity and hierarchical feature mismatch, we introduce a **Multi-Scale Feature Enhancement (MSFE)** module in each decoder stage to refine fused hierarchical features with local, multi-scale, and global context. A **Tri-Attention (TA)** mechanism further refines decoder features, suppressing irrelevant responses while enhancing salient structures and fine-grained boundary cues. Extensive experiments on multiple medical image segmentation benchmarks demonstrate that BiReS consistently outperforms state-of-the-art vision-only and vision-language methods, with notable gains on challenging targets exhibiting substantial scale variation and complex boundary structures.

1 INTRODUCTION

Medical image segmentation stands as one of the most critical and highly challenging tasks in medical image analysis. By automatically delineating the precise location, morphology, and other key characteristics of lesions, it plays an indispensable role in disease identification (Liu et al., 2019), diagnosis (Aggarwal et al., 2021), and prognostic assessment(Tran et al., 2021). While existing deep learning methods (Cao et al., 2022; Chen et al., 2021; Zhou et al., 2018; Isensee et al., 2021), epitomized by the U-Net (Ronneberger et al., 2015) architecture, have achieved remarkable progress, these purely vision-based models are heavily reliant on massive amounts of high-quality pixel-level annotations (Tajbakhsh et al., 2020), and their performance improvements have gradually encountered a bottleneck. Recently, integrating text prompts to assist visual segmentation has emerged as a promising frontier to overcome this limitation. By leveraging the rich semantic priors embedded in clinical reports or descriptive prompts, vision-language segmentation models (Li et al., 2023; Zhong et al., 2023; Bui et al., 2024; Huang et al., 2024; Zhou et al., 2025b) can effectively steer the network's attention toward specific anatomical structures or pathological regions, thereby achieving performance that explicitly surpasses uni-modal approaches.

Despite these advances, text-guided medical image segmentation remains challenging from both visual and cross-modal perspectives . Medical targets exhibit substantial variations in scale, morphology, and appearance, ranging from small localized lesions to large diffuse abnormalities (Wang et al., 2022a; 2024), while low tissue contrast and poorly defined boundaries further complicate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

precise delineation. These visual ambiguities are compounded by a granularity mismatch between global textual semantics and hierarchical pixel-level visual representations. Representative multimodal methods commonly incorporate language through feature fusion or cross-modal attention (Li et al., 2023; Zhong et al., 2023; Bui et al., 2024; Huang et al., 2024), where textual information mainly serves as auxiliary guidance for refining visual features, with relatively limited adaptation of textual representations to image-specific evidence. This asymmetric interaction may restrict the effective exploitation of complementary multimodal information. In addition, hierarchical decoding introduces another source of feature mismatch, as skip-connected encoder features and upsampled deep-level representations differ in both spatial scale and semantic depth. Although their spatial resolutions can be aligned before fusion, their representational characteristics remain heterogeneous. Direct fusion of such heterogeneous features may hinder effective cross-scale integration, particularly for targets exhibiting large scale variations and poorly defined boundaries.

To address these challenges, we propose **BiReS**, a multimodal medical image segmentation framework that couples *bidirectional semantic recalibration* with cross-scale feature refinement. First, we introduce a **Semantic Recalibrator (SR)** to enable reciprocal adaptation between textual and visual representations. In the text-to-vision direction, global textual semantics adaptively modulate visual responses along both spatial and channel dimensions; in the vision-to-text direction, attention-weighted global pooling extracts image-specific visual evidence to update the textual representation. These complementary operations form a closed-loop semantic recalibration process, allowing textual semantics and visual evidence to adapt to each other during segmentation. Second, to mitigate target scale heterogeneity and hierarchical feature mismatch, we introduce a **Multi-Scale Feature Enhancement (MSFE)** module into each decoder block. MSFE refines the fused hierarchical representations by aggregating complementary local, multi-scale, and global context, enabling more effective cross-scale integration and fine-grained enhancement. Furthermore, a lightweight **Tri-Attention (TA)** mechanism selectively refines the decoder features across pixel, channel, and spatial dimensions, suppressing irrelevant responses while preserving boundary-sensitive cues. Multi-level deep supervision is additionally applied to provide direct optimization signals and encourage consistent predictions across scales. Overall, our contributions are summarized as follows:

- We propose **BiReS**, a multimodal medical image segmentation framework that combines bidirectional semantic recalibration with cross-scale feature refinement to better exploit complementary textual and visual information.
- We introduce an **SR** that performs text-to-vision spatial-channel modulation and image-conditioned visual-to-text semantic updating, enabling reciprocal refinement between textual and visual representations.
- We develop a scale-aware decoding strategy that integrates **MSFE**, lightweight **TA**, and multi-level deep supervision to enhance cross-scale representations, suppress irrelevant responses, and preserve fine-grained boundary information.

2 Related Work

Vision-Only Medical Image Segmentation. Deep learning has significantly advanced medical image segmentation through robust feature representation and end-to-end modeling capabilities. Classic CNN-based frameworks, such as U-Net (Ronneberger et al., 2015), V-Net (Milletari et al., 2016), and nnU-Net (Isensee et al., 2021), established the foundation for automatic organ and lesion segmentation by employing encoder-decoder architectures with various skip-connection strategies. To further enhance model representational capacity, AD-LA Former (Wang et al., 2025) combines dynamic convolution with positional attention mechanisms to accommodate the segmentation of complex anatomical structures. nnWNet (Zhou et al., 2025c) integrates convolutional and Transformer representations within the nnU-Net framework to enable continuous propagation and fusion of local and global features. MCADS (Wazir & Kim, 2025) proposes a decoder that combines feature fusion, channel-spatial weighting, and spatial restoration capabilities. Transformer-based models, such as TransUNet (Chen et al., 2021), combine spatial semantic features extracted by CNNs with global contextual information modeled by Transformers to improve performance. Swin-UMamba (Liu et al., 2024) leverages Mamba-based architectures with ImageNet pretraining to efficiently capture long-range dependencies for medical image segmentation. In addition, Zhang et al.

(2026) proposes a decoder enhancement framework that fuses multi-domain features and optimizes skip connections, yielding notable improvements in medical segmentation.

Vision-Language Image Segmentation. Since the breakthrough of vision-language models represented by CLIP (Radford et al., 2021), textual semantics have been widely introduced into image analysis, expanding new supervision paradigms and semantic guidance for segmentation tasks (Xu et al., 2022; Lüddecke & Ecker, 2022). Early works such as BRINet (Hu et al., 2020) explicitly model cross-modal dependencies via bidirectional attention to improve segmentation accuracy. LAVT (Yang et al., 2022) employs an early fusion scheme with language gating to integrate text and Vision Transformer features. In medical image processing, GLoRIA (Huang et al., 2021) learns global-local multimodal representations by aligning image sub-regions with paired report words via contrastive learning. TGANet (Tomar et al., 2022) improves segmentation generalization across varying polyp sizes and quantities using textual embeddings and auxiliary tasks. Similarly, LViT (Li et al., 2023) and Threads (Zhong et al., 2023) employ text-guided strategies that fuse medical text annotations with image features to improve segmentation performance. RecLMIS (Huang et al., 2024) achieves precise visual-textual alignment through a cross-modal conditional reconstruction framework, outperforming existing methods. HCFNet (Zhou et al., 2025a) further enhances feature discriminability using a multi-stage cross-modal contrastive loss. MIT-SAM (Zhou et al., 2025b) integrates SAM (Kirillov et al., 2023) and BERT (Devlin et al., 2019), and employs image-text fusion and self-supervised text reconstruction to capture key semantic cues, such as lesion location and quantity. FMISeg (Yu et al., 2025) exploits language-guided frequency-domain feature interaction to improve visual representation and semantic consistency in medical image segmentation. DD-CMD (Rahman & Hammond, 2026) leverages dual-domain language guidance through spatial cross-attention and frequency-aware modulation for text-guided medical image segmentation, while MedCLIPSeg (Koleilat et al., 2026) adapts CLIP with probabilistic cross-modal attention to achieve data-efficient and uncertainty-aware segmentation.

3 METHOD

3.1 OVERVIEW

As illustrated in Fig. 1, **BiReS** follows an *encoder-bridge-decoder* architecture for text-guided medical image segmentation. The backbone is built upon LViT with PLAM (Li et al., 2023) to extract hierarchical visual representations and incorporate textual cues at early stages. At the two deepest encoder stages, we introduce an **SR** to perform bidirectional semantic refinement between textual and visual representations. The decoder further employs **MSFE** and **TA** to capture scale-aware context and refine features across pixel, channel, and spatial dimensions. Multi-level deep supervision is used during training to provide direct optimization signals to intermediate decoder stages.

Text Encoding & Query Extraction. We use a frozen CLIP text encoder Radford et al. (2021) to obtain global and token-level textual representations. A lightweight Query Extractor summarizes the token features into compact semantic queries, which are injected into the LViT backbone for early-stage multimodal fusion. The global text embedding is further used by SR for text-to-vision modulation and vision-to-text semantic updating. Implementation details are provided in Appendix A.

3.2 BIDIRECTIONAL SEMANTIC RECALIBRATION

To alleviate the granularity mismatch between textual and visual representations (Ding et al., 2021; Liu et al., 2023) and the asymmetry of text-to-visual guidance (Wang et al., 2022b), SR couples text-conditioned visual modulation with image-conditioned semantic updating. Specifically, it leverages text semantics to recalibrate visual features while simultaneously enriching textual representations with image-specific visual context. Considering the varying semantic granularity and abstraction levels across hierarchical features, SR is embedded at specific stages of the network, passing both refined representations to the next layer within each stack. For a single SR layer, let $\mathbf{V} \in \mathbb{R}^{C\times N}$ denote the current visual feature after spatial flattening, where $C = C_i$ and $N = H_iW_i$, and let $\mathbf{s} \in \mathbb{R}^{d_s}$ denote the current semantic state. The stage and layer indices are omitted for clarity.

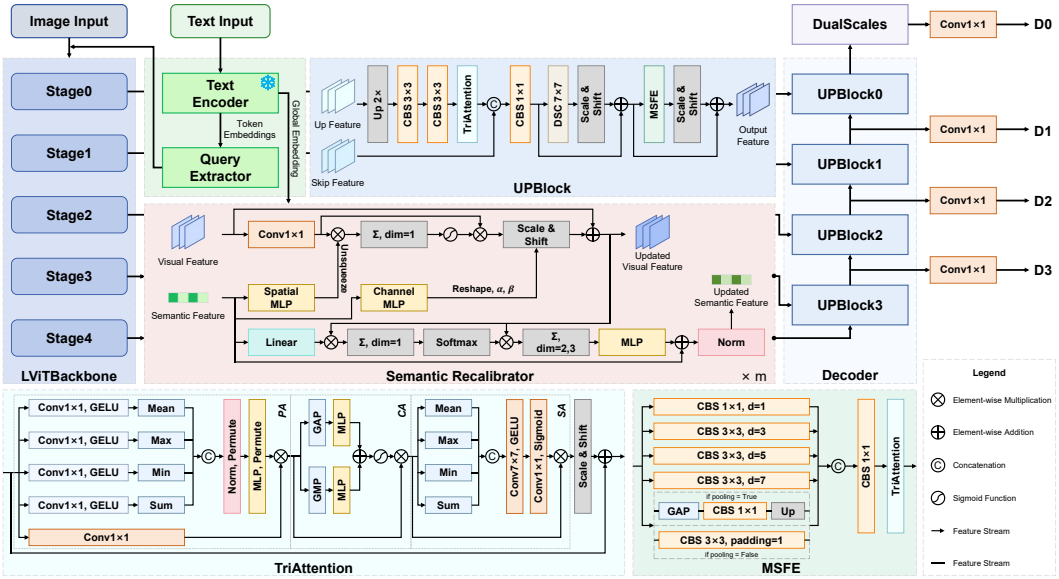

Figure 1: Overview of the proposed BiReS architecture. CBS denotes a sequential block of convolution, batch normalization, and SiLU activation, and DSC represents depthwise separable convolution. The LViT backbone extracts hierarchical features at five resolutions. The deepest features are refined by SR for bidirectional semantic recalibration, while MSFE and TA enhance multi-scale decoder representations.

Text-to-Vision Recalibration. A 1×1 convolution f_v projects the visual feature $\mathbf{V}$ into $\mathbf{X} = f_v(\mathbf{V}) \in \mathbb{R}^{C \times N}$. Two separate two-layer MLPs with GELU activation map $\mathbf{s}$ to a spatial weighting vector $\mathbf{u} = \mathrm{MLP}_s(\mathbf{s}) \in \mathbb{R}^C$ and channel modulation parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \mathrm{MLP}_c(\mathbf{s})$, with $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^C$. Spatial gating and FiLM-style channel modulation (Perez et al., 2017) are then combined as

$$\begin{aligned} \mathbf{a} &= \sigma(\mathbf{X}^\top \mathbf{u}), \\ \mathbf{V}^+ &= \mathbf{V} + (\mathbf{X} \odot \mathbf{a}) \odot (\mathbf{1} + \boldsymbol{\gamma}) + \boldsymbol{\beta}, \end{aligned} \tag{1}$$

where σ denotes the sigmoid function, $\mathbf{a} \in \mathbb{R}^N$ is broadcast across channels, and $\odot$ denotes element-wise multiplication with broadcasting. The spatial gate selects text-relevant image regions, while the channel-wise scale and shift adapt their feature responses. The residual path retains the incoming visual representation.

Visual-to-Text Updating. To adapt the text semantic to image-specific evidence, a linear projection $f_s : \mathbb{R}^{d_s} \to \mathbb{R}^C$ maps the current text embedding to a visual query. Its similarity to the refined visual features determines attention weights for global pooling:

$$\begin{aligned} \mathbf{p} &= \mathrm{softmax}\big((\mathbf{V}^+)^\top f_s(\mathbf{s})\big), \\ \mathbf{s}^+ &= \mathrm{LN}\big(\mathbf{s} + \mathrm{MLP}_{v2s}(\mathbf{V}^+ \mathbf{p})\big)\,. \end{aligned} \tag{2}$$

Here, the softmax normalizes over the N spatial positions, $\mathbf{V}^+\mathbf{p} \in \mathbb{R}^C$ is the attention-pooled visual context, MLP_{v2s} maps this context to $\mathbb{R}^{d_s}$, and LN denotes layer normalization. The updated text semantic $\mathbf{s}^+$ conditions the next SR layer, forming an iterative exchange between textual semantics and visual evidence across the stack. The refined visual features are reshaped to their original spatial layouts for decoding.

3.3 Tri-Attention

TA sequentially applies pixel, channel, and spatial attention to suppress irrelevant responses and enhance salient, boundary-sensitive features. Given an input feature $\mathbf{U} \in \mathbb{R}^{C \times H \times W}$, the output is

$$\mathrm{TA}(\mathbf{U}) = \mathbf{U} + \mathrm{LS}(\mathcal{A}_{\mathrm{sa}}(\mathcal{A}_{\mathrm{ca}}(\mathcal{A}_{\mathrm{pa}}(\mathbf{U})))\,, \tag{3}$$

where $\mathcal{A}_{\mathrm{pa}}$, $\mathcal{A}_{\mathrm{ca}}$, and $\mathcal{A}_{\mathrm{sa}}$ denote the pixel, channel, and spatial attention operators, respectively, and LS is a learnable channel-wise affine transformation.

Pixel Attention. Pixel attention constructs local descriptors from four separately projected feature maps. Let $\mathcal{K} = \{\mathrm{avg}, \max, \min, \mathrm{sum}\}$, and let ρ_k reduce a feature map along its channel axis using the corresponding statistic. With independent 1×1 convolution f_k, pixel attention is defined as:

$$\begin{aligned} \mathbf{R}(\mathbf{U}) &= \mathrm{BN}([\rho_k(\mathrm{GELU}(f_k(\mathbf{U})))]_{k\in\mathcal{K}}), \\ \mathcal{A}_{\mathrm{pa}}(\mathbf{U}) &= \mathrm{MLP}(\mathbf{R}(\mathbf{U})) \odot f_p(\mathbf{U}), \end{aligned} \tag{4}$$

where $\mathbf{R}(\mathbf{U}) \in \mathbb{R}^{H\times W\times 4}$ denotes the normalized multi-statistic descriptor, BN denotes batch normalization, and f_p is another 1×1 convolution. The two-layer MLP transforms $\mathbf{R}(\mathbf{U})$ into multiplicative attention weights, which are applied element-wise to $f_p(\mathbf{U})$.

Channel & Spatial Attention. Channel attention adaptively reweights feature channels according to their relative importance, while spatial attention models cross-channel dependencies at each spatial location. Following the dual-pooling design of CBAM (Woo et al., 2018), channel attention aggregates spatial information using global average pooling (GAP) and global max pooling (GMP). The two descriptors pass through separate MLPs before being combined into channel weights. Spatial attention instead summarizes the input across channels as $\mathbf{S}(\mathbf{U}) = [\rho_k(\mathbf{U})]_{k\in\mathcal{K}} \in \mathbb{R}^{4\times H\times W}$ and maps this descriptor to a single spatial gate:

$$\begin{aligned} \mathcal{A}_{\mathrm{ca}}(\mathbf{U}) &= \mathbf{U} \odot \sigma(\mathrm{MLP}_{\mathrm{avg}}(\mathrm{GAP}(\mathbf{U})) + \mathrm{MLP}_{\max}(\mathrm{GMP}(\mathbf{U}))), \\ \mathcal{A}_{\mathrm{sa}}(\mathbf{U}) &= \mathbf{U} \odot \sigma(g_{1\times1}(\mathrm{GELU}(g_{7\times7}(\mathbf{S}(\mathbf{U}))))). \end{aligned} \tag{5}$$

where $g_{7\times7}$ and $g_{1\times1}$ are convolutions that preserve the spatial resolution. Each operator receives the preceding operator's output in Eq. equation 3. Unlike pixel attention, which computes statistics after separate learned projections, spatial attention computes its statistics directly from its incoming features.

3.4 Multi-Scale Feature Enhancement

Spatially aligned decoder and skip features can still differ in semantic content and receptive field. MSFE refines their fused representation by combining local detail, dilated context, and global information. Given a fused feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, MSFE uses five parallel branches, each producing $C_{\mathrm{mid}} = C/4$ channels. All convolutions within these branches and their fusion layer use a Convolution-BatchNorm-SiLU (CBS) unit.

The local branch $\mathcal{B}_1$ applies a 1×1 CBS unit to retain spatial detail. Three contextual branches $\mathcal{B}_2, \mathcal{B}_3, \mathcal{B}_4$ apply 3×3 CBS units with dilation rates 3, 5, and 7, respectively, using matching padding to preserve the spatial resolution. The global branch $\mathcal{B}_5$ applies global average pooling, a 1×1 CBS unit, and bilinear upsampling to $H \times W$. At the decoder stage where global pooling is disabled, this branch instead uses a resolution-preserving 3×3 CBS unit.

The branch outputs are concatenated, compressed by a 1×1 CBS unit, and refined by TA:

$$\begin{aligned} \mathbf{Z} &= \mathrm{CBS}_{1\times1}\big([\mathcal{B}_1(\mathbf{X}), \ldots, \mathcal{B}_5(\mathbf{X})]\big), \\ \mathrm{MSFE}(\mathbf{X}) &= \mathrm{TA}(\mathbf{Z}). \end{aligned} \tag{6}$$

This aggregation exposes the decoder to complementary receptive fields, while the subsequent Tri-Attention operation selectively refines the fused responses by enhancing informative responses and suppressing redundant activations. The modular design of MSFE facilitates integration into diverse CNN backbones with minimal architectural modifications.

3.5 Multi-Level Deep Supervision

To stabilize training dynamics and improve semantic consistency across different scales, we adopt a four-level deep supervision strategy. We attach a 1×1 convolutional prediction head to each of the four decoder levels. The highest-resolution feature additionally passes through the *dual-scale refinement module* (Appendix B) before its prediction head. Let $\widehat{\mathbf{Y}}_i$ denote the prediction at level

i and $\mathbf{Y}_i$ the ground-truth mask matched to its spatial resolution, with $i = 0$ denoting the highest-resolution level. Training minimizes

$$\mathcal{L}_{\text{total}} = \sum_{i=0}^{3} \alpha_i \left[\mathcal{L}_{\text{Dice}}(\widehat{\mathbf{Y}}_i, \mathbf{Y}_i) + \mathcal{L}_{\text{CE}}(\widehat{\mathbf{Y}}_i, \mathbf{Y}_i) \right], \tag{7}$$

where α_i weights the contribution of each supervision level, and $\mathcal{L}_{\text{Dice}}$ and $\mathcal{L}_{\text{CE}}$ denote the Dice and cross-entropy losses, respectively. At inference, only the highest-resolution prediction $\widehat{\mathbf{Y}}_0$ is retained.

4 Experiments

4.1 Experimental Setup

Datasets & Evaluation Metrics. We evaluate BiReS on three public medical image segmentation datasets: QaTa-COV19 (Degerli et al., 2022), MosMedData+ (Morozov et al., 2020; Hofmanninger et al., 2020), and BUSI (Al-Dhabyani et al., 2020). As these datasets do not provide paired text annotations, we adopt the annotations introduced by prior vision-language segmentation methods, using LViT Li et al. (2023) for QaTa-COV19 and MosMedData+, and MedCLIPSeg (Koleilat et al., 2026) for BUSI. Dataset splits follow the corresponding protocols, and normal BUSI samples without lesions are excluded since our study focuses on lesion segmentation. We evaluate segmentation performance using four standard metrics: Dice, IoU, HD95, and ASD. Dice and IoU quantify region-level overlap, while HD95 and ASD measure boundary consistency. As the datasets are provided in PNG format without physical spacing metadata, HD95 and ASD are computed in the 2D image coordinate space, with distances reported in pixels. All metrics are computed per sample and macro-averaged across the dataset to ensure equal contribution from each image.

Implementation Details. Our model is implemented using PyTorch 2.12.0, and all experiments are conducted on an NVIDIA RTX 4090 GPU. We employ the LViT encoder as the backbone network. During the training phase, we utilize Automatic Mixed Precision (AMP) with a fixed batch size of 32. The network is optimized using the AdamW optimizer with an initial learning rate of 4×10^{-4} and a weight decay of 1×10^{-2}. For learning rate scheduling, we adopt the CosineAnnealingWarmRestarts scheduler, setting T_0, T_{mult}, and η_{min} to 20, 1, and 1×10^{-5}, respectively. The maximum number of training epochs is set to 300, and an early stopping mechanism is applied with a patience of 40 epochs.

Baselines. We compare BiReS with a diverse set of segmentation baselines, including four unimodal methods: UNet (Ronneberger et al., 2015), TransUNet (Chen et al., 2021), Swin-UMamba (Liu et al., 2024), and nnWNet (Zhou et al., 2025c), as well as six multimodal methods: LViT (Li et al., 2023), RecLMIS (Huang et al., 2024), Thread (Zhong et al., 2023), MMI-UNet (Bui et al., 2024), DD-CMD (Rahman & Hammond, 2026), and MedCLIPSeg (Koleilat et al., 2026). We reproduce all baselines using their official implementations and evaluate them under the same experimental protocol, including identical hardware settings, text annotations, dataset splits, training iterations, and evaluation metrics.

4.2 Comparison with State-of-the-Art Methods

We compare BiReS with existing state-of-the-art methods on QaTa-COV19, MosMedData+, and BUSI, considering region overlap (Dice and IoU), boundary accuracy (HD95 and ASD), and the contribution of textual guidance. As shown in Tables 1 and 2, BiReS consistently achieves superior performance across all datasets, demonstrating its effectiveness in multimodal medical image segmentation.

Region and Boundary Accuracy. BiReS achieves the highest Dice and IoU scores on all three datasets. Specifically, it improves over the strongest overlap baselines by 0.67/0.79 pp, 1.01/0.39 pp, and 1.49/1.27 pp in Dice/IoU on QaTa-COV19, MosMedData+, and BUSI, respectively. Meanwhile, BiReS obtains the lowest HD95 and ASD values, indicating more accurate boundary localization. These results demonstrate that BiReS improves both region-level segmentation and contour delineation.

Table 1: Region overlap performance on QaTa-COV19, MosMedData+, and BUSI. Dice and IoU scores are reported as percentages (%). $\mathrm{BiReS}_{\mathrm{w/o\ text}}$ refers to the text-free variant of BiReS.

Methods	Text	Params (M)	QaTa-COV19		MosMedData+		BUSI	
			Dice ↑	IoU ↑	Dice ↑	IoU ↑	Dice ↑	IoU ↑
UNet	✗	31.0	80.30	71.12	71.42	57.89	75.63	67.14
TransUNet	✗	105.3	80.72	71.50	70.42	56.37	80.49	72.01
Swin-UMamba	✗	59.9	79.92	70.64	71.42	57.81	79.40	70.50
nnWNet	✗	7.0	79.45	69.86	70.82	56.73	80.31	72.28
LViT	✓	38.9	83.36	74.64	73.14	59.89	79.97	72.03
RecLMIS	✓	69.4	83.97	75.52	75.09	62.56	80.12	72.11
Thread	✓	44.0	84.44	75.88	72.35	58.95	82.01	74.35
MMI-UNet	✓	56.2	81.13	71.69	61.93	48.11	77.63	69.14
DD-CMD	✓	47.8	84.67	76.35	71.80	58.22	81.45	73.20
MedCLIPSeg	✓	18.7	83.64	74.76	56.28	42.36	81.39	72.63
$\mathrm{BiReS}_{\mathrm{w/o\ text}}$	✗	24.6	80.81	71.76	72.70	59.08	80.74	72.83
BiReS (Ours)	✓	64.3	**85.34**	**77.14**	**76.10**	**62.95**	**83.50**	**75.62**

Table 2: Boundary delineation accuracy on QaTa-COV19, MosMedData+, and BUSI. HD95 and ASD scores are reported in pixels.

Methods	Text	Params (M)	QaTa-COV19		MosMedData+		BUSI	
			HD95 ↓	ASD ↓	HD95 ↓	ASD ↓	HD95 ↓	ASD ↓
UNet	✗	31.0	29.87	8.30	24.46	5.22	31.28	12.30
TransUNet	✗	105.3	29.96	8.21	23.73	4.89	23.38	9.30
Swin-UMamba	✗	59.9	30.85	8.76	21.64	4.19	23.68	9.66
nnWNet	✗	7.0	31.67	8.30	28.84	5.32	22.91	8.88
LViT	✓	38.9	18.22	4.76	19.84	4.61	29.68	10.62
RecLMIS	✓	69.4	16.05	4.08	19.70	4.89	27.89	10.38
Thread	✓	44.0	16.62	4.19	19.86	4.49	21.02	8.46
MMI-UNet	✓	56.2	18.90	4.81	26.32	4.64	22.09	8.63
DD-CMD	✓	47.8	16.16	4.09	19.35	3.98	20.95	8.00
MedCLIPSeg	✓	18.7	16.49	4.52	33.27	8.60	20.21	7.81
$\mathrm{BiReS}_{\mathrm{w/o\ text}}$	✗	24.6	29.25	8.01	24.87	4.44	21.55	8.77
BiReS (Ours)	✓	64.3	**14.79**	**3.61**	**18.87**	**3.36**	**18.97**	**7.69**

Robustness to Complex Lesions. MMI-UNet and MedCLIPSeg show notable performance degradation on MosMedData+, where lesions often exhibit irregular shapes, complex boundaries, and disconnected regions, suggesting that these approaches may have limitations in modeling complex lesion structures. In contrast, BiReS maintains stable performance on MosMedData+, demonstrating stronger robustness and generalization ability under challenging segmentation scenarios.

Effect of Textual Guidance. Multimodal methods consistently outperform vision-only baselines across all three datasets, supporting the complementary role of textual guidance. Compared with $\mathrm{BiReS}_{\mathrm{w/o\ text}}$, the full model improves Dice by 4.53, 3.40, and 2.76 pp on QaTa-COV19, MosMedData+, and BUSI, respectively, with consistent gains in IoU, HD95, and ASD. These results highlight the complementary role of textual cues alongside visual features in improving both region prediction and boundary localization.

4.3 Ablation Study

Effectiveness of Different Components. Table 3(a) evaluates SR, MSFE, and TA by removing each component individually and all three jointly. Removing SR causes the largest boundary degradation among the individual ablations, increasing HD95/ASD from 14.79/3.61 to 16.34/4.07 pixels.

Removing MSFE mainly affects region overlap and HD95, reducing Dice/IoU by 0.33/0.43 pp and increasing HD95 by 0.37 pixels, while ASD remains nearly unchanged. In contrast, removing TA yields the largest overlap degradation, with Dice/IoU decreasing by 0.87/0.94 pp. Removing all three components further reduces Dice to 84.02% and increases HD95/ASD to 18.22/4.61 pixels. These results demonstrate that SR, MSFE, and TA provide complementary benefits within the full architecture, while their ablation effects should not be interpreted as independent or additive contributions.

Deep-Supervision Weighting. Table 3(b) compares four fixed deep-supervision weighting schemes. The geometrically decreasing schedule, $\alpha_i = 2^{-i}$, assigns the largest weight to the highest-resolution prediction while retaining auxiliary supervision at coarser scales. This schedule achieves the best performance across all four metrics. Compared with supervising only the highest-resolution output, it improves Dice and IoU by 0.68 and 0.85 pp, respectively, while reducing HD95 and ASD by 1.06 and 0.24 pixels. It also slightly outperforms uniform weighting, with gains of 0.11/0.10 pp in Dice/IoU and reductions of 0.33/0.01 pixels in HD95/ASD. Notably, reversing the schedule while preserving the same total weight decreases Dice from 85.34% to 85.07% and increases HD95 from 14.79 to 15.59 pixels. These results suggest that assigning greater supervision weight to finer-resolution predictions is more effective in this setting. We therefore adopt the fixed geometrically decreasing schedule. In preliminary experiments, we also considered learnable supervision weights but observed inferior performance. One possible explanation is that adaptive weighting may overemphasize auxiliary objectives that are easier to optimize, thereby reducing the optimization pressure on the final high-resolution prediction.

Table 3: Component ablations and deep-supervision schedule abaltions on QaTa-COV19. Weight vectors are ordered from the highest to the lowest prediction resolution.

(a) Component ablations

Method	Dice (%)↑	IoU (%)↑	HD95↓	ASD↓
BiReS (Ours)	**85.34**	**77.14**	**14.79**	**3.61**
w/o SR	84.85	76.65	16.34	4.07
w/o MSFE	85.01	76.71	15.16	3.62
w/o TA	84.47	76.20	15.50	3.78
w/o ALL	84.02	75.69	18.22	4.61

(b) Deep-supervision weights

Weights	Dice (%)↑	IoU (%)↑	HD95↓	ASD↓
[1.0, 0.0, 0.0, 0.0]	84.66	76.29	15.85	3.85
[1.0, 1.0, 1.0, 1.0]	85.23	77.04	15.12	3.62
[0.125, 0.25, 0.5, 1.0]	85.07	76.75	15.59	3.82
[1.0, 0.5, 0.25, 0.125]	**85.34**	**77.14**	**14.79**	**3.61**

SR Placement & Stack Depth. As shown in Table 4(a), we first evaluate the placement of a single SR layer across encoder levels. The stronger results are generally observed at deeper levels: Level 3 performs best on Dice, IoU, and HD95, while Level 4 achieves the lowest ASD. Although no single placement is optimal across all metrics, the overall trend suggests that semantic recalibration is more effective when applied to higher-level features. We further examine SR stack depth at Level 4, as reported in Table 4(b). A two-layer stack provides the best overall trade-off, improving overlap accuracy and HD95 over a single layer with only a slight increase in ASD. Increasing the depth to three layers degrades both overlap and boundary metrics, indicating that additional recalibration depth does not provide a monotonic benefit. These results favor a shallow SR stack at deeper encoder levels, while the conclusions are limited to the separate placement and Level-4 depth sweeps rather than a joint search over both factors.

Table 4: Ablation study of SR configurations on QaTa-COV19.

(a) Insertion level; stack depth = 1

Level	Dice (%)↑	IoU (%)↑	HD95↓	ASD↓
Level 0	84.77	76.42	16.19	4.02
Level 1	84.66	76.39	16.02	4.19
Level 2	84.91	76.66	15.68	3.96
Level 3	**85.15**	**76.93**	**14.84**	3.74
Level 4	85.13	76.84	15.08	**3.50**

(b) Stack depth; insertion level = 4

SR layers	Dice (%)↑	IoU (%)↑	HD95↓	ASD↓
1	85.13	76.84	15.08	**3.50**
2	**85.23**	**77.01**	**14.77**	3.57
3	85.06	76.78	15.42	3.83

Effect of Bidirectional Semantic Recalibration. To examine the role of vision-to-text feedback, we compare the full SR module with the w/o SR baseline and a T2V-only variant that retains text-conditioned spatial and channel modu-

Table 5: Ablation of SR recalibration.

Variant	T2V	V2T	Dice (%)↑	IoU (%)↑	HD95↓	ASD↓
w/o SR	✗	✗	84.85	76.65	16.34	4.07
T2V-only	✓	✗	84.88	76.74	15.29	3.77
Full SR	✓	✓	**85.34**	**77.14**	**14.79**	**3.61**

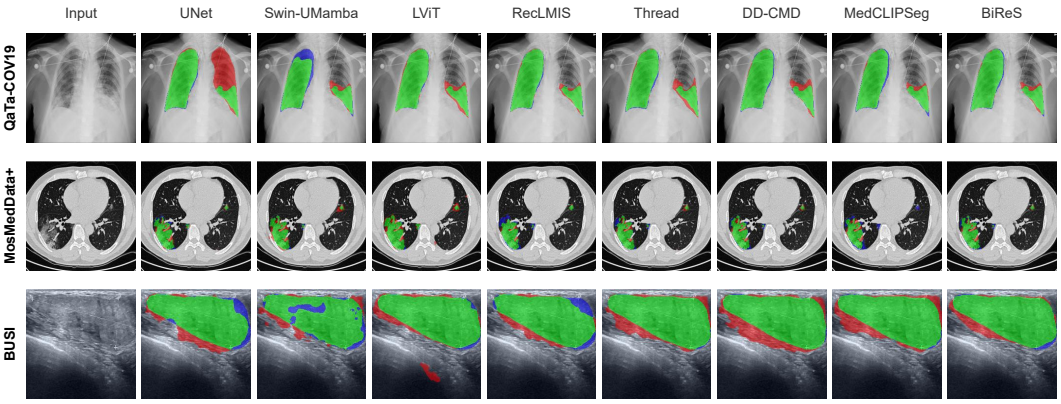

Figure 2: Qualitative comparison of selected methods on QaTa-COV19, MosMedData+, and BUSI. TP, FP, and FN are highlighted in green, red, and blue, respectively.

lation while disabling vision-to-text updating. As shown in Table 5, the bidirectional design consistently improves both overlap and boundary metrics over the T2V-only variant. Notably, T2V-only achieves overlap performance similar to w/o SR but substantially better HD95 and ASD, suggesting that semantic conditioning primarily improves sensitivity to lesion boundaries. Further enabling vision-to-text updating provides additional gains, indicating that image-specific evidence in the textual representation complements text-conditioned visual modulation. Overall, these results support bidirectional semantic recalibration for improving segmentation and boundary localization.

4.4 Qualitative Analysis

Fig. 2 presents qualitative comparisons across different models. BiReS produces fewer false-positive regions and more accurate lesion delineation across the shown examples. In the BUSI case, its predicted contour most closely matches the ground-truth boundary.

Effect of Semantic Recalibration. To examine how SR affects hierarchical visual representations, we visualize intermediate segmentation probability maps from the corresponding decoder stages with and without SR in Fig. 3, covering large lesions, small lesions, and complex boundaries. Without SR, coarse-stage responses are more diffuse and show weaker boundary awareness across the three cases. For small lesions, incorporating SR yields stronger responses in the target region, consistent with the positional cues provided by the text prompt. For complex boundaries, the model without SR largely misses the lesion on the left, whereas SR produces a clear response.

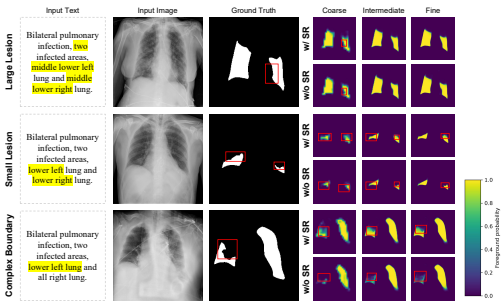

Figure 3: Effect of SR on hierarchical decoder predictions.

5 Conclusion

In this paper, we presented **BiReS**, a multimodal medical image segmentation framework that combines bidirectional semantic recalibration with scale-aware feature refinement. The proposed SR enabled reciprocal interaction between textual and visual representations, while MSFE and Tri-Attention improved multi-scale contextual modeling and fine-grained structural refinement. Experiments on QaTa, MosMed, and BUSI demonstrated consistent improvements in both overlap-based segmentation performance and surface-based boundary accuracy over strong vision-only and vision-language baselines. Future work will investigate more flexible text conditioning and extend BiReS to broader imaging modalities and clinical segmentation scenarios.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and presentation, including polishing and condensing parts of the manuscript (e.g., the method section) and improving the consistency of mathematical notation and formula presentation. We also used generative AI tools to assist in searching for and identifying recent literature and methods published in 2026 that were potentially relevant to our work and subsequent comparative experiments.

We did not rely on generative AI tools to formulate the core research hypotheses, develop the proposed methodology, derive mathematical claims or proofs, or interpret the experimental results. All literature and methods suggested through AI-assisted searches were manually verified by the authors against the original publications before being included or used in our experiments. The comparative experiments and their reported results were conducted and checked by the authors. All AI-assisted revisions to the manuscript were also reviewed by the authors to ensure technical correctness and consistency with the intended meaning.

We take full responsibility for the final content of this work, including all text, claims, experimental results, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of all experimental results reported in this work. The implementation of the proposed method and all comparative experiments is fully reproducible from our codebase. The main paper and appendix provide the necessary details of the proposed method, experimental protocols, datasets, evaluation metrics, and implementation settings. Upon completion of the review process, we will publicly release the complete source code on GitHub, together with all training and evaluation configurations, including hyperparameter settings, model configurations, data-processing scripts, and experiment scripts required to reproduce the reported results.

REFERENCES

Ravi Aggarwal, Viknesh Sounderajah, Guy Martin, Daniel SW Ting, Alan Karthikesalingam, Dominic King, Hutan Ashrafian, and Ara Darzi. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine*, 4(1):65, 2021.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Phuoc-Nguyen Bui, Duc-Tai Le, and Hyunseung Choo. Visual-textual matching attention for lesion segmentation in chest images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 702–711. Springer, 2024.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022.

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

Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2306–2310. IEEE, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16321–16330, 2021.

Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European radiology experimental*, 4(1):50, 2020.

Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4424–4433, 2020.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3942–3951, 2021.

Xiaoshuang Huang, Hongxiang Li, Meng Cao, Long Chen, Chenyu You, and Dong An. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(4):1821–1835, 2024.

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

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

Taha Koleilat, Hojat Asgariandehkordi, Omid Nejatimanzari, Berardino Barile, Yiming Xiao, and Hassan Rivaz. Medclipseg: probabilistic vision-language adaptation for data-efficient and generalizable medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1406–1417, 2026.

Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 43(1):96–107, 2023.

Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18653–18663, 2023.

Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International conference on medical image computing and computer-assisted intervention*, pp. 615–625. Springer, 2024.

Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7086–7096, 2022.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Ieee, 2016.

Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, AV Vladzymyrskyy, Natalya V Ledikhova, Victor A Gombolevskiy, Ivan A Blokhin, Pavel B Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. URL https://arxiv.org/abs/1709.07871.

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

Md Maklachur Rahman and Tracy Hammond. Dual-domain cross-modal decoding for clinical text-guided medical image segmentation. *arXiv preprint arXiv:2608.11335*, 2026.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical image analysis*, 63:101693, 2020.

Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. Tganet: Text-guided attention for improved polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 151–160. Springer, 2022.

Khoa A Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D Williams, John V Pearson, and Nicola Waddell. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome medicine*, 13(1):152, 2021.

Yan Wang, Yangqin Feng, Lei Zhang, Joey Tianyi Zhou, Yong Liu, Rick Siow Mong Goh, and Liangli Zhen. Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images. *Medical Image Analysis*, 81:102535, 2022a.

Yan Wang, Liangli Zhen, Tien-En Tan, Huazhu Fu, Yangqin Feng, Zizhou Wang, Xinxing Xu, Rick Siow Mong Goh, Yipin Ng, Claire Calhoun, et al. Geometric correspondence-based multimodal learning for ophthalmic image analysis. *IEEE Transactions on Medical Imaging*, 43(5):1945–1957, 2024.

Yuenan Wang, Hua Wang, and Fan Zhang. A medical image segmentation model with auto-dynamic convolution and location attention mechanism. *Computer Methods and Programs in Biomedicine*, 261:108593, 2025.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11686–11695, 2022b.

Saad Wazir and Daeyoung Kim. Rethinking decoder design: Improving biomarker segmentation using depth-to-space restoration and residual linear attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30861–30871, 2025.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18134–18144, 2022.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18155–18165, 2022.

Bo Yu, Jianhua Yang, Zetao Du, Yan Huang, Chenglong Li, and Liang Wang. Frequency-domain multi-modal fusion for language-guided medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 278–288. Springer, 2025.

Fan Zhang, Zhiwei Gu, and Hua Wang. Decoding with structured awareness: integrating directional, frequency-spatial, and structural attention for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12421–12429, 2026.

Yi Zhong, Mengqiu Xu, Kongming Liang, Kaixin Chen, and Ming Wu. Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 724–733. Springer, 2023.

Xichuan Zhou, Qianqian Song, Jing Nie, Yujie Feng, Haijun Liu, Fu Liang, Lihui Chen, and Jin Xie. Hybrid cross-modality fusion network for medical image segmentation with contrastive learning. *Engineering Applications of Artificial Intelligence*, 144:110073, 2025a.

Xichuan Zhou, Lingfeng Yan, Rui Ding, Chukwuemeka Clinton Atabansi, Jing Nie, Lihui Chen, Yujie Feng, and Haijun Liu. Mit-sam: Medical image-text sam with mutually enhanced heterogeneous features fusion for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2025b.

Yanfeng Zhou, Lingrui Li, Le Lu, and Minfeng Xu. nnwnet: Rethinking the use of transformers in biomedical image segmentation and calling for a unified evaluation benchmark. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20852–20862. IEEE, 2025c.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pp. 3–11. Springer, 2018.

A Text Encoding and Query Extraction

We adopt the frozen text encoder of the pre-trained CLIP Radford et al. (2021) to extract textual representations. Given a tokenized text prompt T, the encoder outputs two types of features: a global text embedding $\mathbf{s} \in \mathbb{R}^d$, and a sequence of hidden states $\mathbf{H} \in \mathbb{R}^{L\times d}$, where L and d denote the sequence length and feature dimension, respectively. To distill task-relevant semantic information and construct structured semantic inputs for visual interaction, we design a **Query Extractor** module. Specifically, we initialize N_q learnable query embeddings $\mathbf{Q} \in \mathbb{R}^{N_q\times d}$. A cross-attention mechanism is then applied to aggregate contextual information, utilizing $\mathbf{Q}$ as the queries and $\mathbf{H}$ as both keys and values:

$$\mathbf{T}_{\text{token}} = \text{CrossAttn}(\mathbf{Q}, \mathbf{H}, \mathbf{H}) \tag{8}$$

The resulting compact semantic queries $\mathbf{T}_{\text{token}} \in \mathbb{R}^{N_q\times d}$ are fed into the **LViTBackbone** together with image features for early-stage multimodal fusion, generating a hierarchy of multi-scale features $\{\mathbf{F}_i\}_{i=0}^4$, where $\mathbf{F}_i \in \mathbb{R}^{C_i\times \frac{H}{2^i}\times \frac{W}{2^i}}$. Here, C_i denotes the channel dimension at stage i, while H and W represent the height and width of the input image, respectively.

B Implementation Details of DualScales

We provide the implementation details of **DualScales** in this section. The definitions of Conv-BN-SiLU (CBS) and Tri-Attention (TA) are given in the main text and are therefore omitted here.

Given an input feature map

$$\mathbf{X} \in \mathbb{R}^{C_{\text{in}}\times H\times W}, \tag{9}$$

we set the intermediate channel dimension to

$$C_m = \left\lfloor \frac{C_{\text{in}}}{2} \right\rfloor. \tag{10}$$

DualScales processes $\mathbf{X}$ through two parallel branches:

$$\mathbf{X}_1 = \text{CBS}_{1\times1}(\mathbf{X}), \qquad \mathbf{X}_2 = \text{CBS}_{3\times3}(\mathbf{X}), \tag{11}$$

where both branches produce C_m output channels. The 3×3 operator uses stride 1, padding 1, and dilation 1, such that the spatial resolution is preserved.

The two branch outputs are concatenated along the channel dimension and subsequently fused using a 1×1 CBS block:

$$\mathbf{X}_f = \mathrm{CBS}_{1\times1}\left(\mathrm{Concat}\left[\mathbf{X}_1, \mathbf{X}_2\right]\right), \tag{12}$$

where

$$\mathbf{X}_f \in \mathbb{R}^{C_{\mathrm{out}} \times H \times W}. \tag{13}$$

The fused representation is further processed by the TA module:

$$\mathbf{X}_a = \mathrm{TA}(\mathbf{X}_f). \tag{14}$$

A residual pathway is employed to preserve the input information. Specifically, the shortcut is defined as

$$\mathcal{S}(\mathbf{X}) = \begin{cases} \mathbf{X}, & C_{\mathrm{in}} = C_{\mathrm{out}}, \\ \mathrm{BN}\left(\mathrm{Conv}_{1\times1}(\mathbf{X})\right), & C_{\mathrm{in}} \neq C_{\mathrm{out}}. \end{cases} \tag{15}$$

The final output is computed as

$$\mathbf{Y} = \mathbf{X}_a + \mathcal{S}(\mathbf{X}). \tag{16}$$

Therefore, DualScales can be compactly formulated as

$$\begin{aligned} \mathbf{Y} = \mathrm{TA}\Big(\mathrm{CBS}_{1\times1}\big(\mathrm{Concat}[\mathrm{CBS}_{1\times1}(\mathbf{X}), \\ \mathrm{CBS}_{3\times3}(\mathbf{X})]\big)\Big) + \mathcal{S}(\mathbf{X}). \end{aligned} \tag{17}$$

All convolutional operations in the two branches and the fusion layer use stride 1; hence, DualScales preserves the spatial resolution of the input feature map.