Uncertainty-Aware Vision-Language Segmentation for Medical Imaging

Aryan Das*
VIT Bhopal
aryan.das2021@vitbhopal.ac.in

Tanishq Rachamalla*
SAHE, Andhra Pradesh
tanishqrachamalla12@gmail.com

Koushik Biswas
IIIT Delhi
koushikb@iiitd.ac.in

Swalpa Kumar Roy
Tezpur University Assam
swalpa@tezu.ernet.in

Vinay Kumar Verma
IIT Kanpur
vinayugc@gmail.com

Abstract

We introduce a novel uncertainty-aware multimodal segmentation framework that leverages both radiological images and associated clinical text for precise medical diagnosis. We propose a Modality Decoding Attention Block (MoDAB) with a lightweight State Space Mixer (SSMix) to enable efficient cross-modal fusion and long-range dependency modelling. To guide learning under ambiguity, we propose the Spectral-Entropic Uncertainty (SEU) Loss, which jointly captures spatial overlap, spectral consistency, and predictive uncertainty in a unified objective. In complex clinical circumstances with poor image quality, this formulation improves model reliability. Extensive experiments on various publicly available medical datasets, QATA-COVID19, MosMed++, and Kvasir-SEG, demonstrate that our method achieves superior segmentation performance while being significantly more computationally efficient than existing State-of-the-Art (SoTA) approaches. Our results highlight the importance of incorporating uncertainty modelling and structured modality alignment in vision-language medical segmentation tasks. Code: https://github.com/arya-domain/UA-VLS

1. Introduction

Medical image segmentation is a foundational task in computer-aided diagnosis, surgical planning, and clinical research [1, 2]. Deep learning has enabled automated image segmentation for assessing disease severity and guiding treatment. However, various unimodal methods depend heavily on extensive labelled data, which is often limited in clinical settings [3, 4]. To overcome this, recent studies have explored multimodal segmentation by integrating image data with textual reports. Leveraging natural language as auxiliary supervision offers rich contextual cues, enhancing segmentation performance, especially when visual quality is poor or annotations are sparse.

Vision-language segmentation (VLS) aims to utilize natural language inputs, such as radiology reports or anatomical queries, to guide the segmentation process [5]. This multimodal paradigm offers several advantages: it mitigates the semantic disconnect between low-level visual cues and high-level clinical concepts, reduces the need for task-specific supervision, and enables more intuitive medical workflows [2, 6, 7].

Despite progress in VLS, most existing methods neglect the role of uncertainty modelling during training, which is critical in clinical applications where predictions must be both accurate and reliable. Uncertainty-aware guidance can help models focus on ambiguous regions and reduce overconfident errors, especially when dealing with noisy data. However, uncertainty has largely been explored in unimodal medical segmentation, with minimal adoption in multimodal vision-language frameworks. Furthermore, effective alignment between visual features and language cues remains challenging, often limiting the benefits of cross-modal learning with limited parameters in a model. To address these issues, we incorporate an uncertainty-aware optimization and propose a state-space-based modality [8] integration strategy. This allows for efficient global dependency modelling while keeping the computational cost significantly lower than conventional transformer-based designs. Our contributions can be summarized as follows:

- We propose Modality Decoding Attention Block (MoDAB) and State Space Mixer (SSMix) to enable structured multimodal fusion with long-range dependency modeling for medical vision-language tasks.
- We also introduce Spectral-Entropic Uncertainty (SEU) Loss, a unified objective that integrates spatial, spectral, and uncertainty guidance into a single optimization.
- Our computationally efficient model outperforms the recent State-of-The-Art (SoTA) methods on multiple

[*] Equal Contribution

benchmarks.

2. Related Work

Unimodal Segmentation Models: Early deep learning-based medical image segmentation models were largely built on fully convolutional networks, with U-Net [1] being the most influential. Enhanced variants like UNet++ [3], Attention U-Net [2], and nnUNet [9] improved feature fusion via skip connections, dense pathways, and attention mechanisms. To capture global context, hybrid models emerged: TransUNet [10] combined CNNs with Vision Transformers (ViTs), and Swin-UNet [4] adopted hierarchical Swin Transformer blocks for multi-resolution processing. UCTransNet [11] further improved semantic understanding by integrating cross-fusion transformers and multi-head attention in skip connections. Sequence modelling approaches like U-Mamba [12] and Swin-UMamba [13] introduced Mamba-based modules into the U-Net, enabling long-range spatial dependency modelling through recurrent dynamics as an alternative to attention mechanisms.

State Space Models: State Space Models (SSMs) have emerged as promising alternatives to transformer-based architectures for long-sequence modeling due to their linear time complexity and memory efficiency. Gu et al. [14] proposed S4, a structured state-space sequence model capable of capturing long-range dependencies while remaining computationally efficient. Subsequent advancements, including Hyena [15] and FlashAttention-2 [16], further demonstrated the effectiveness of structured memory mechanisms in sequence learning. Recently, Mamba [8] introduced selective state-space updates, enabling linear-time inference and training for long-range tasks with minimal compute overhead. While SSMs have shown strong results in language and vision domains, their application in multimodal and medical segmentation tasks remains limited.

Vision-Language Segmentation Models: VLS has emerged as a transformative paradigm, enabling models to integrate clinical semantics with spatial reasoning for more interpretable and context-aware predictions. Foundational works like ConVIRT [17], GLoRIA [18], CLIP [5], and BiomedCLIP [19] leveraged contrastive learning on paired medical images and textual reports, producing powerful joint embeddings that served as a backbone for a variety of downstream tasks. These models primarily focused on aligning vision and language representations at a global or hierarchical level, which laid the groundwork for segmentation models that could benefit from such multimodal understanding. Building upon these pretrained foundations, transformer-based models such as ViLT [6], LAVT [20], and LViT-T [21] introduced mechanisms to directly inject textual information into the visual encoding pipeline using cross-modal attention, enabling dense prediction models to utilize linguistic prompts describing lesions, anatomical regions, or disease types. CMIRNet [22] advanced this paradigm by introducing sophisticated cross-modal interactive reasoning mechanisms specifically designed for referring image segmentation in medical contexts. Meanwhile, architectures like Ariadne [7] and SLViT [23] leveraged report-based supervision and multimodal attention to bridge the semantic gap in segmentation settings, using text as surrogate annotations to improve spatial localization.

More recent and specialized frameworks introduced tighter coupling between the two modalities; for example, TMCA [24] incorporated contrastive objectives at multiple levels of the network to improve feature alignment and semantic understanding across both modalities. Similarly, RecLMIS [25] employed a training paradigm where each modality helped reconstruct the other, encouraging shared latent understanding of anatomical and contextual features. MulModSeg [26] addressed the challenging problem of unpaired multi-modal medical image segmentation by introducing modality-conditioned text embedding and alternating training strategies. Likewise, DMMI [27] introduced dual-memory structures to separately capture and interact with visual and textual cues, reinforcing consistency and context awareness. Structured learning approaches such as RefSegformer [28] and TGANet [29] incorporated external references and graph-based language guidance, using textual anchors or graph attention mechanisms to disambiguate complex visual regions. Models like LGA [30] and TMC [31] pushed the limits of token-level and hierarchical language conditioning, embedding semantic meaning directly into the encoding and decoding stages of segmentation. Anatomical Structure-Guided Medical Vision-Language Pre-training represents a significant advancement in foundation model development by incorporating explicit anatomical structure awareness into the pre-training process. Finally, scalable models such as MAdapter [32] proposed language-guided adapters that could be inserted into vision transformers, enabling efficient multimodal learning without retraining the entire backbone.

Uncertainty in Medical Segmentation: Uncertainty estimation has emerged as a critical component in medical image segmentation, particularly for enhancing model reliability in high-stakes clinical environments. Zeevi et al. [33] introduced Monte-Carlo Frequency Dropout (MC-FD), a technique that extends traditional MC-Dropout to the frequency domain. Their method demonstrated improved calibration and delineation of boundaries across diverse modalities, including MRI and CT. Similarly, Antico et al. [34] performed a comprehensive evaluation of uncertainty quantification techniques across multiple algorithms and datasets, concluding that pixel-wise uncertainty estimation, especially using MC-Dropout, significantly improves the robustness and interpretability of segmentation models. Entropy is one of the most interpretable and computationally efficient measures of uncertainty. Sedai et al. [35] utilized pixel-wise entropy maps to estimate aleatoric uncertainty in retinal vessel segmentation. Similarly, Roy et al. [36] demonstrated the use of entropy from softmax outputs to capture both model and data uncertainty. These works highlight how entropy-based uncertainty can improve the interpretability of predictions and flag ambiguous regions, which is essential for downstream tasks.

3. Methodology

3.1. Modalities Encoding

We utilize two pre-trained models to encode the input modalities: *ConvNeXt-Tiny* [37] as the visual encoder and *BioViL CXR-BERT* [38] as the text encoder. The visual encoder, denoted as $\mathscr{V}_{\mathscr{E}}$, extracts hierarchical features from four stages, capturing both fine-grained and abstract semantic information. Given a batch of chest X-ray images $\mathscr{I} \in \mathbb{R}^{B\times 3\times H\times W}$ where B is the batch size, H is height and W is the width, the visual encoder outputs a set of multi-scale feature maps:

$$\mathscr{I}'_i = \mathscr{V}_{\mathscr{E}}(\mathscr{I}) \tag{1}$$

where $\mathscr{I}'_i \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$ denotes the feature map extracted at stage $i \in \{1,2,3,4\}$. These features are spatially aligned and serve as inputs to subsequent modality decoding attention blocks.

For the textual input, we utilize a frozen text encoder, denoted as $\mathscr{T}_{\mathscr{E}}$, to extract contextualized token embeddings. Let the input token sequence be represented as $\mathscr{T} = [t_1, t_2, \ldots, t_N]$, where N denotes the sequence length. The encoder outputs a sequence of semantic embeddings:

$$\mathscr{T}' = \mathscr{T}_{\mathscr{E}}(\mathscr{T}) \tag{2}$$

where $\mathscr{T}' \in \mathbb{R}^{B\times N\times D}$ denotes the output feature matrix, where each token is embedded in a D-dimensional space. These embeddings capture rich contextual semantics essential for cross-modal alignment.

3.2. Modality Decoding Attention Block (MoDAB)

The *Modality Decoding Attention Block (MoDAB)* fuses spatial visual representations with contextual textual embeddings through a series of operations: Multi-Head Self-Attention ($SelfAttn$), Cross-Attention ($CrossAttn$) with Sinusoidal Positional Encodings (SPE), and a State Space Mixer (SSMix), which is a sequence mixer (detailed in Section 3.3).

Let $\mathbf{X} \in \mathscr{I}'_i$ denote the visual input from the i^{th} stage of the visual encoder. The textual input $\mathscr{T}'$ is first projected to match the visual space via a learnable transformation, followed by a state-space-based mixer:

$$\mathscr{T}_{SSMix} = GELU\left(SSMix\left(LeakyReLU\left(Linear(\mathscr{T}')\right)\right)\right) \tag{3}$$

where $\mathscr{T}_{SSMix} \in \mathbb{R}^{B\times N\times Y_i}$, B is the batch size and $Y_i = H_i \times W_i$ is the projected dimension.

Self-Attention: We apply *Multi-Head Self Attention (MHSA)* to the visual sequence $\mathbf{X}$ to capture intra-modal dependencies among spatial tokens. First, the input is normalized (LN) and augmented with Sinusoidal Positional Encodings (SPE):

$$\mathbf{X}' = SPE\left(LN(\mathbf{X})\right) \tag{4}$$

where $\mathbf{X}' \in \mathbb{R}^{B\times C_i\times Y_i}$ is linearly projected into h attention heads, each with dimension D_k, using learned weight matrices:

$$\mathbf{Q}_{SA_j} = \mathbf{X}'\mathbf{W}^{Q}_{SA_j}, \quad \mathbf{K}_{SA_j} = \mathbf{X}'\mathbf{W}^{K}_{SA_j}, \quad \mathbf{V}_{SA_j} = \mathbf{X}'\mathbf{W}^{V}_{SA_j} \tag{5}$$

where $\mathbf{W}^{Q}_{SA_j}, \mathbf{W}^{K}_{SA_j}, \mathbf{W}^{V}_{SA_j} \in \mathbb{R}^{C_i\times D_k}$ for each head $j \in \{1,\ldots,h\}$.

The attention for each head is computed using the scaled dot-product formulation:

$$head_j = SoftMax\left(\frac{\mathbf{Q}_{SA_j}\mathbf{K}^{\top}_{SA_j}}{\sqrt{D_k}}\right)\mathbf{V}_{SA_j} \tag{6}$$

All heads are concatenated and passed through a final projection:

$$SelfAttn(\mathbf{X}') = Concat(head_1, \ldots, head_h)\mathbf{W}^{O} \tag{7}$$

where $\mathbf{W}^{O} \in \mathbb{R}^{h\cdot D_k\times C}$ is a learned projection matrix. No masking is applied since all spatial tokens attend to each other. The MHSA output is normalized and added residually to form the self-attended feature:

$$\mathbf{X_{SA}} = \mathbf{X}' + LN\left(SelfAttn(\mathbf{X}')\right) \tag{8}$$

where $\mathbf{X_{SA}} \in \mathbb{R}^{B\times C_i\times Y_i}$.

Cross-Attention: The *Multi-Head Cross-Attention (MHCA)* extends MHSA by enabling cross-modal interaction: the query (Q) is derived from one modality, while the key (K) and value (V) come from another. Here, the self-attended visual features $\mathbf{X}_{SA}$ act as the query, and the state-space-enhanced textual embeddings $\mathscr{T}_{SSMix}$ provide the key and value. Similar to MHSA, both inputs are normalized and augmented with SPE to retain spatial and sequential structure.

$$\mathbf{Q}_{CA_j} = SPE\left(LN(\mathbf{X}_{SA})\right) \tag{9}$$

$$\mathbf{K}_{CA_j} = SPE\left(\mathscr{T}_{SSMix}\right), \quad \mathbf{V}_{CA_j} = \mathscr{T}_{SSMix} \tag{10}$$

The attention mechanism computes relevance between the visual queries and textual keys, producing cross-attended visual representations:

$$\widehat{\mathbf{X}}_{\mathbf{CA}} = CrossAttn(\mathbf{Q}_{CA_j}, \mathbf{K}_{CA_j}, \mathbf{V}_{CA_j}) \tag{11}$$

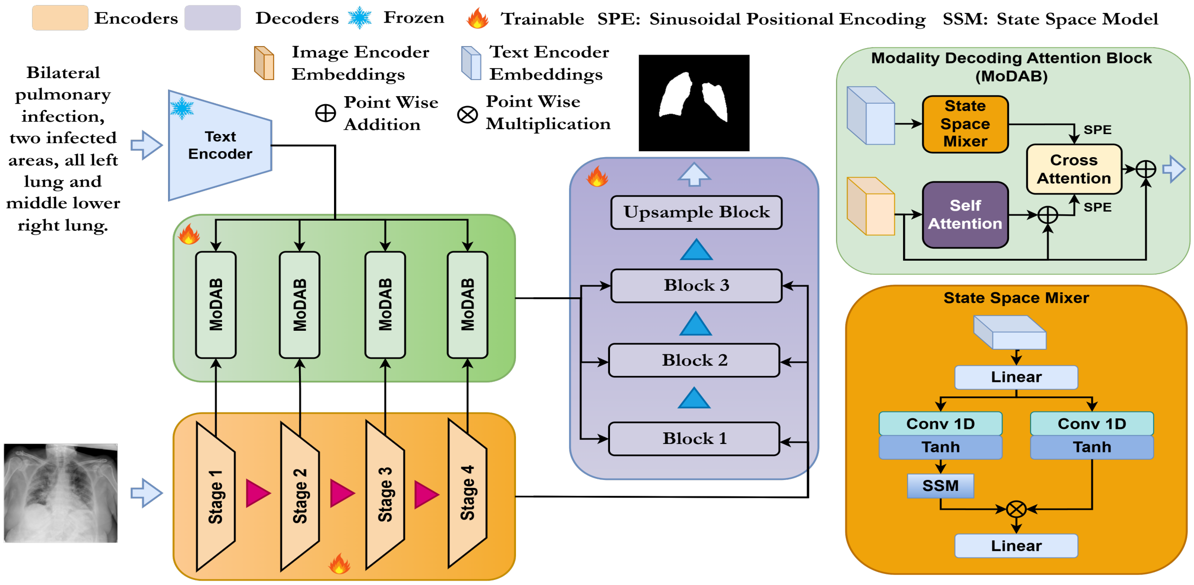

Figure 1. Overview of the proposed architecture. The model integrates visual and frozen text encoders and the Modality Decoding Attention Block (MoDAB), which incorporates Self-Attention and Cross-Attention along with a State Space Mixer (SSMix) for efficient multimodal fusion. The decoder reconstructs segmentation masks from the fused features through a multi-stage upsampling pathway.

To enable adaptive integration of textual context, the cross-attention output is normalized and added to the original visual features, scaled by a learnable scalar parameter $\alpha \in \mathbb{R}$:

$$\mathbf{F} = \mathbf{X} + \alpha \cdot LN\left(\widehat{\mathbf{X}}_{\mathbf{CA}}\right) \tag{12}$$

where $\alpha \in \mathbb{R}$ is randomly initialized and learned during training. $\mathbf{F} \in \mathbb{R}^{B \times C_i \times Y_i}$ captures both spatial visual dependencies and semantically aligned textual cues. This enriched feature map is subsequently propagated to the decoder for segmentation mask reconstruction.

3.3. State Space Mixer (SSMix)

The *State Space Mixer (SSMix)* is designed to enhance long-range dependency modelling in sequential data by combining learned temporal dynamics with convolutional operations and selective scanning mechanisms. Additionally, it incorporates a gating technique, serving as an efficient and lightweight module. Given the textual input $\mathscr{T}' \in \mathbb{R}^{B \times N \times D}$, the module outputs a transformed feature matrix $\mathscr{T}_{SSMix} \in \mathbb{R}^{B \times N \times Y_i}$. The input $\mathscr{T}'$ is first projected into an intermediate representation of size $2 \cdot d_{\text{inner}}$, where $d_{\text{inner}} = \gamma D$ and γ is the expansion factor:

$$\mathscr{T}_H = Linear(\mathscr{T}') \in \mathbb{R}^{B \times N \times 2d_{\text{inner}}} \tag{13}$$

The projected features are then transposed to prepare for 1D convolution and split along the channel dimension into two parts:

$$\mathbf{P}, \mathbf{Q} = Split\left(Transpose(\mathscr{T}_H, [0, 2, 1])\right), \tag{14}$$

where $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{B \times d_{\text{inner}} \times N}$.

After splitting the features, depthwise 1D convolutions are applied to each component to extract localized temporal features:

$$\tilde{\mathbf{P}} = tanh(Conv1D_x(\mathbf{P})), \quad \tilde{\mathbf{Q}} = tanh(Conv1D_z(\mathbf{Q})) \tag{15}$$

The output $\tilde{\mathbf{P}}$ is passed through a linear projection to produce dynamic time-step parameters Δ and state parameters $\mathbf{B}, \mathbf{C}$:

$$[\Delta, \mathbf{B}, \mathbf{C}] = Split\left(Linear(\tilde{\mathbf{P}})\right) \tag{16}$$

The stepping weights $\Delta \in \mathbb{R}^{B \times d_{inner} \times L}$ are further refined through a softplus reparameterization of their log-transformed initializations to ensure stability:

$$\Delta = Softplus(\Delta + \text{bias}_{\Delta}) \tag{17}$$

A state-space update is then performed using the *selective State Space Model (SSM)*, which models latent dynamics across time steps with exponentially decaying memory kernels. The state output **SCAN** is computed as:

$$\mathbf{SCAN} = SSM(\tilde{\mathbf{X}}, \Delta, \mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{E}) \tag{18}$$

where **E** is a learned gating vector applied to modulate the scan dynamics and **A** is a diagonal state transition matrix.

Finally, the output **SCAN** is concatenated with the convolutional branch $\tilde{\mathbf{Q}}$, and the result is projected back to the output embedding dimension using a final linear transformation:

$$\mathscr{T}_{SSMix} = Linear\left(Concat(\mathbf{SCAN}, \tilde{\mathbf{Q}})^{\top}\right) \tag{19}$$

The resulting $\mathscr{T}_{SSMix}$ captures both global and local dependencies, facilitating effective multimodal fusion in downstream decoding.

3.4. Decoder

The decoder reconstructs the spatial segmentation layout by first reshaping the fused multimodal feature $\mathbf{F} \in \mathbb{R}^{B \times C_i \times Y_i}$ into a spatial feature map $\mathbf{F}' \in \mathbb{R}^{B \times C_i \times H_i \times W_i}$. It follows a four-stage decoding pipeline that progressively restores the spatial resolution. For each stage $m \in \{1, 2, 3\}$, an *Upsampling Block* doubles the spatial resolution using a transposed convolution operation:

$$\mathbf{F}_{\text{up}}^{(m)} = \text{TransConv}(\mathbf{F}^{(m-1)}), \quad \mathbf{F}^{(0)} := \mathbf{F}' \tag{20}$$

where $\text{TransConv}(\cdot)$ denotes a 2×2 transposed convolution with stride 2. The upsampled feature map $\mathbf{F}_{\text{up}}^{(m)} \in \mathbb{R}^{B \times C_m \times H_m \times W_m}$ captures progressively finer spatial structure, with C_m representing the output channels at stage m.

The upsampled feature $\mathbf{F}_{\text{up}}^{(m)}$ is concatenated with the corresponding encoder feature $\mathscr{I}'_{4-m}$ at the same resolution level. The resulting tensor is processed by a *Convolutional Refinement Block (CRB) CRB$_m$*, comprising two convolutional layers, LeakyReLU activations, and batch normalization:

$$\mathbf{F}_{\text{CRB}}^{m} = CRB_m\left(Concat(\mathbf{F}_{\text{up}}^{(m)}, \mathscr{I}'_{4-m})\right) \tag{21}$$

The final stage applies a *Subpixel Upsampling Network (SUN)*, consisting of a convolutional layer followed by pixel shuffling. The convolution increases the feature dimensionality:

$$\mathbf{F}_{\text{pre}} = Conv2D(\mathbf{F}_{\text{CRB}}^{3}) \tag{22}$$

Pixel shuffling $\Pi(\cdot)$ rearranges spatial elements to produce a high-resolution output by a factor of r in each spatial dimension:

$$\mathbf{F}_{\text{SU}} = \Pi(\mathbf{F}_{\text{pre}}) \tag{23}$$

yielding $\mathbf{F}_{\text{SU}} \in \mathbb{R}^{B \times C \times rH \times rW}$.

To improve local consistency and mitigate boundary artifacts, an average pooling operation with $o \times o$ kernel and appropriate padding is applied:

$$\mathbf{F}_{\text{avg}} = AvgPool2D(Pad(\mathbf{F}_{\text{SU}})) \tag{24}$$

Finally, a 1×1 convolutional output layer maps the refined features into the desired number of prediction channels:

$$\hat{\mathbf{Y}} = Conv_{1 \times 1}(\mathbf{F}_{\text{avg}}) \in \mathbb{R}^{B \times C_o \times H \times W} \tag{25}$$

where C_o is the number of output channels. This multi-stage decoding process enables coarse-to-fine segmentation reconstruction, preserving both semantic and spatial detail through visual-textual alignment.

3.5. Objective Function

To guide the model toward anatomically precise, structurally consistent, and uncertainty-aware predictions, we introduce the *Spectral-Entropic Uncertainty (SEU) Loss*, a unified objective designed for medical vision-language segmentation. Rather than treating separate objectives independently, SEU Loss holistically integrates spatial, spectral, and probabilistic priors into a single formulation.

Let $\hat{\mathbf{Y}} \in \mathbb{R}^{B \times C \times H \times W}$ denote the predicted segmentation map and $\hat{\mathbf{G}} \in \mathbb{R}^{B \times C \times H \times W}$ the one-hot encoded ground truth. The SEU loss is expressed as:

$$\begin{aligned} \mathscr{L}\text{SEU} = \mathscr{L}\text{Dice}(\hat{\mathbf{Y}}, \hat{\mathbf{G}}) + \lambda_{\text{F}} \cdot \mathscr{R}\text{Spectral}(\hat{\mathbf{Y}}, \hat{\mathbf{G}}) \\ + \lambda_{\text{E}} \cdot \mathscr{R}\text{Entropy}(\hat{\mathbf{Y}}) \end{aligned} \tag{26}$$

where λ_{F} and λ_{E} are modulation weights for spectral alignment and uncertainty regularization, respectively. Each component contributes to a different representational aspect, but collectively they form a single landscape.

Spatial Alignment: The core supervision comes from a differentiable Dice loss, capturing the pixel-level overlap between $\hat{\mathbf{Y}}$ and $\hat{\mathbf{G}}$:

$$\mathscr{L}_{\text{Dice}} = 1 - \frac{2 \cdot \sum(\hat{\mathbf{Y}} \cdot \hat{\mathbf{G}}) + \varepsilon}{\sum \hat{\mathbf{Y}} + \sum \hat{\mathbf{G}} + \varepsilon} \tag{27}$$

where the summation is over all spatial and channel dimensions, and ε is a small constant for numerical stability.

Spectral Consistency: To enforce global structural fidelity, we align the magnitude of Fourier spectra between the predicted and target masks:

$$\mathscr{R}_{\text{Spectral}} = \left|\left|\mathscr{F}(\hat{\mathbf{Y}})\right| - \left|\mathscr{F}(\hat{\mathbf{G}})\right|\right|_2^2 \tag{28}$$

where $\mathscr{F}(\cdot)$ denotes the 2D Fourier Transform and $|\cdot|$ is the magnitude operation. This encourages preservation of global anatomical topology, especially beneficial for diffuse or subtle lesions.

Uncertainty Guidance: To penalize ambiguous predictions and promote confident outputs, we incorporate an entropy-based regularization term defined as:

$$\mathscr{R}_{\text{entropy}} = -\frac{1}{BHW} \sum_{b,c,h,w} \hat{\mathbf{Y}}_{b,c,h,w} \log(\hat{\mathbf{Y}}_{b,c,h,w} + \delta) \tag{29}$$

where the indices b, c, h, and w range over $b \in \{1, \ldots, B\}$, $c \in \{1, \ldots, C\}$, $h \in \{1, \ldots, H\}$, and $w \in \{1, \ldots, W\}$, corresponding to the batch size, number of classes, and spatial dimensions (height and width), respectively. The term δ is a small constant added for numerical stability to prevent undefined values. This entropy-based regularization term acts as a soft constraint, encouraging the model to reduce uncertainty by promoting low-entropy, confident predictions.

4. Experiments

4.1. Experimental Setup

All experiments were carried out on a high-performance computing server equipped with an *Intel Xeon Silver 4214R*

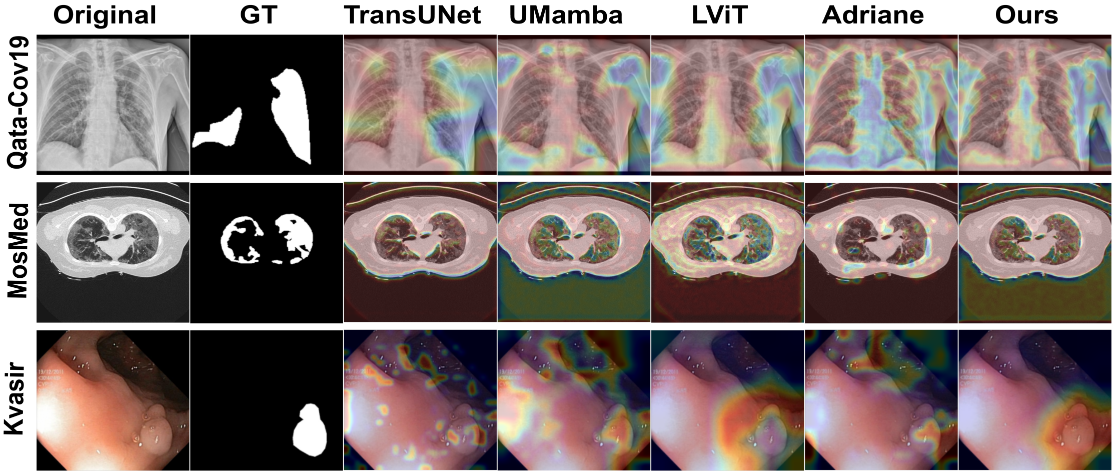

Figure 2. Comparison of Grad-CAM-Based Attention Visualizations Between the Proposed Model and Baseline methods

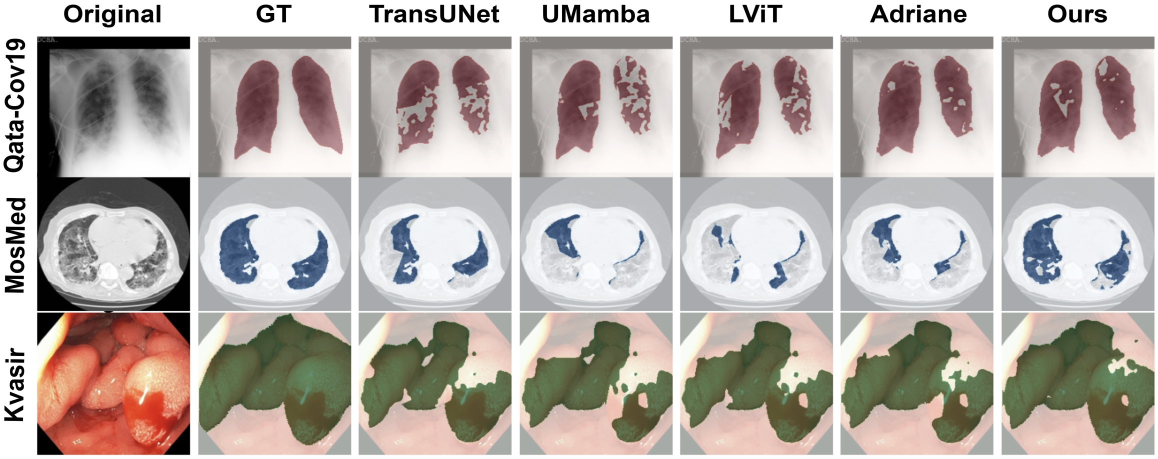

Figure 3. Qualitative Comparison of Predicted Segmentation Maps with Baseline Models

CPU running at 2.40 GHz, 128 GB of RAM, and *NVIDIA A30* GPUs. The system environment was configured with *CUDA version 12.2* to ensure GPU acceleration for training and evaluation.

4.2. Datasets

We employ three publicly available datasets: **QaTa-COV19**, **MosMed++**, and **Kvasir-SEG** to evaluate the proposed model. These datasets, initially developed for unimodal segmentation purposes, have been augmented with concise natural language descriptions by recent works such as LViT [21] and MedVLSM [39], enabling vision-language segmentation.

QaTa-COV19: The QaTa-COV19 dataset [40], compiled by researchers from Qatar University and Tampere University, comprises 9,258 chest X-ray images of COVID-19 cases. It is one of the first datasets to include manually annotated COVID-19 lesion regions. To support multimodal training, this dataset was extended with textual descriptions by LViT [21].

MosMed++: MosMed++ [41, 42] is a chest CT dataset containing 2,729 axial slices from patients diagnosed with COVID-19. Each slice is annotated with severity scores and enriched with textual descriptions curated by LViT [21].

Kvasir-SEG: The Kvasir-SEG dataset [43] includes 1,000 high-resolution gastrointestinal endoscopy images with pixel-wise polyp annotations. The image resolutions vary from 332×487 to 1920×1072 pixels. MedVLSM [39] introduced caption annotations describing polyp characteristics. For consistency with other datasets, we selected a single caption per image, prioritizing those that included location-based descriptors.

4.3. Training Details

Experiments were conducted using a consistent training and validation setup across all three datasets. Input images were uniformly resized to a resolution of 224×224 pixels to ensure compatibility with the model architecture and to standardize training across datasets, thereby maintaining consistency in spatial features. A batch size of 32 was used dur-

Table 1. Comparison of Monomodal and Multimodal State-of-The-Art (SoTA) methods on medical image segmentation across three datasets: QaTa-COV19, MosMed++, and Kvasir-SEG. Metrics include Dice score (%), mean Intersection over Union (mIoU, %), number of Trainable Parameters (Millions), and Floating-Point Operations (FLOPs) per second (Billions). **Black**: best, **Green**: second best, **Blue**: third best values.

Modality	Method	Trainable Params (M)	Flops (G)	QATA-COV19		MosMedData++		Kvasir-Seg	
				Dice (%)	mIoU (%)	Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
MonoModels	U-Net [1]	14.8	50.3	78.91	69.32	64.30	50.50	82.33	74.26
	UNet++ [3]	74.5	94.6	79.47	70.05	71.63	58.14	82.79	73.94
	AttUNet [2]	34.9	101.9	79.11	69.83	66.07	52.68	82.94	74.17
	nnUNet [9]	19.1	412.7	80.30	70.62	72.32	60.14	84.22	75.41
	TransUNet [10]	105	56.7	78.44	68.84	71.13	58.28	90.53	85.94
	Swin-Unet [4]	82.3	67.3	77.85	68.07	63.19	49.93	89.09	85.84
	UCTransNet [11]	65.6	63.2	79.00	69.34	65.71	52.55	91.04	87.31
	Swin-UMamba [13]	60	68	80.02	70.11	65.31	51.28	79.43	68.63
	U-Mamba [12]	18.51	375.78	80.51	70.89	65.88	52.17	89.81	85.14
Multimodal	ConVIRT [17]	35.2	44.6	79.45	70.29	71.92	59.52	89.24	83.01
	TGANet [29]	**19.8**	41.9	79.66	70.61	71.63	59.00	89.76	83.20
	CLIP [5]	87	105.3	79.57	70.54	71.75	59.43	90.04	86.29
	BiomedClip [19]	87	105.3	87.75	78.10	66.33	50.27	85.34	77.60
	GLoRIA [18]	45.6	60.8	79.83	70.54	72.36	60.12	86.10	77.93
	ViLT [6]	87.4	55.9	79.45	70.02	72.10	59.85	86.42	76.91
	LAVT [20]	118.6	83.8	79.15	69.73	73.10	60.14	87.16	74.72
	LViT [21]	29.7	54.1	83.40	74.89	74.32	61.03	87.59	75.16
	Ariadne [7]	43.95	**22.36**	**88.06**	**79.00**	**78.29**	64.44	90.45	82.34
	SLViT [23]	131.5	51.1	79.10	68.71	72.36	60.55	89.88	83.03
	DMMI [27]	114.6	63.3	83.85	75.42	74.78	61.59	**90.96**	83.41
	RefSegformer [28]	195	103.6	83.93	75.34	74.81	61.46	90.57	**83.69**
	RecLMIS [25]	**23.7**	**24.1**	84.93	76.86	77.26	**64.95**	86.58	77.08
	LGA [30]	**8.24**	382.17	84.40	76.05	62.30	**75.43**	89.82	83.25
	MAdapter [32]	-	-	**90.07**	**81.88**	**78.40**	62.77	**91.37**	**84.36**
	Our Model	39.9	**17.87**	**92.24**	**84.9**	**79.67**	**66.38**	**93.83**	**87.62**

ing both training and validation phases. The training process was executed for a maximum of 200 epochs, with early stopping implemented based on a patience of 20 epochs to avoid overfitting and ensure generalizability. A minimum training duration was enforced, requiring at least 20 epochs to promote model stability during the initial learning phase. The values of λ_F and λ_E are 0.3 and 0.1, respectively.

We also employed the AdamW optimizer [44], which decouples weight decay from the gradient update process, making it particularly effective for transformer-based architectures. The initial learning rate was set to 5×10^{-4} for the MosMed++ dataset and 3×10^{-4} for the QaTa-COV19 and Kvasir-SEG datasets, based on preliminary tuning and empirical observations for optimal convergence. Additionally, a cosine annealing learning rate scheduler [45] was utilized to progressively reduce the learning rate over time, with a maximum cycle length of 200 epochs and a minimum learning rate threshold of 1×10^{-6}.

5. Results and Discussion

5.1. Qualitative and Quantitative Analysis

As shown in Figure 2, our model exhibits more focused and semantically aligned attention compared to SoTAs. In Figure 3, we present qualitative segmentation results on three datasets, highlighting only the main predicted regions for clarity. Our model demonstrates superior precision in localizing and delineating the target areas compared to SoTAs.

Our proposed model demonstrates superior performance across all datasets, achieving state-of-the-art results in both Dice coefficient and mean Intersection over Union (mIoU) metrics (Table 1). On the QATA-COV19 dataset, the top-performing baselines include MAdapter (90.07% Dice, 81.88% mIoU), BiomedClip (87.75% Dice, 78.10% mIoU), and RecLMIS (84.93% Dice, 76.86% mIoU) among multimodal approaches, while U-Mamba (80.51% Dice, 70.89% mIoU) and nnUNet (80.30% Dice, 70.62% mIoU) represent the best monomodal methods. Our model achieved exceptional performance with a Dice score of 92.24% and mIoU of 84.9%, demonstrating +2.17% improvement over MAdapter, +4.49% over BiomedClip, and **+11.73%** over the best monomodal approach, U-Mamba.

On the MosMedData++ dataset, the top-performing baselines include MAdapter (78.40% Dice, 62.77% mIoU), Ariadne (78.29% Dice, 64.44% mIoU), and RecLMIS (77.26% Dice, 64.95% mIoU) among multimodal approaches, while nnUNet (72.32% Dice, 60.14% mIoU) and UNet++ (71.63% Dice, 58.14% mIoU) represent the best monomodal methods. Our model achieved a Dice score of 79.67% and mIoU of 66.38%, establishing new state-of-the-art results with +1.27% improvement over MAdapter, +1.38% over Ariadne, and +7.35% over the best monomodal method, nnUNet.

On the Kvasir-Seg dataset for polyp segmentation, the top-performing baselines include MAdapter (91.37% Dice, 84.36% mIoU), UCTransNet (91.04% Dice, 87.31% mIoU), and DMMI (90.96% Dice, 83.41% mIoU) among multimodal approaches, while TransUNet (90.53% Dice, 85.94% mIoU) and U-Mamba (89.81% Dice, 85.14% mIoU) represent the best monomodal methods. Our model achieved outstanding performance with a Dice score of 93.83% and mIoU of 87.62%, demonstrating +2.46% improvement over MAdapter, +2.79% over UCTransNet, and +3.3% over the best monomodal approach, TransUNet.

5.2. Computational Efficiency Analysis

Our proposed model demonstrates remarkable computational efficiency while maintaining superior performance, as shown in Table 1. With only 39.9M trainable parameters, our model is significantly more compact than many SoTAs, such as RefSegformer (195M), SLViT (131.5M), and LAVT (118.6M). Our model requires only 17.87G FLOPs, making it the most efficient in the SoTAs. Despite being more efficient than most baselines, our model consistently achieves the highest performance across all datasets, demonstrating an excellent performance-efficiency trade-off.

6. Ablation Studies

To assess the contributions of key components in our framework, we conduct a detailed ablation study using the Kvasir-SEG dataset. The experiments are categorized into three aspects: Loss formulation, Textual Guidance, and Architectural Components. Results in Table 2 highlight the performance drop associated with the removal or replacement of specific modules, thereby validating their necessity.

Loss Function Analysis: We evaluate the impact of our proposed Spectral-Entropic Uncertainty (SEU) loss by replacing it with commonly used alternatives. Replacing SEU with Dice loss or binary cross-entropy (BCE) results in noticeable performance degradation (Dice: 93.44% and 92.03%, respectively), confirming SEU's advantage in capturing both spatial and uncertainty-aware features.

Table 2. Ablation results on the **Kvasir-SEG** dataset. Each condition evaluates the model after removing or replacing a key component. Metrics reported are Dice score (%) and mIoU (%).

Method	**Dice (%)**	**mIoU (%)**
Loss Function		
Dice loss	93.44	85.76
BCE loss	92.03	85.26
Textual Guidance		
Inference w/o Text Prompts	87.28	81.32
Training w/o MoDAB	85.15	73.86
Architectural Replacements		
SSMix with Linear layer	91.72	82.43
Cross-Attention with Addition	92.11	82.59
Complete Model (ours)	**93.86**	**87.62**

Effect of Textual Guidance: The role of vision-language alignment is examined by removing text prompts during inference and training. Omitting textual inputs during inference causes the Dice score to drop to 87.28%. Eliminating textual supervision entirely by removing the MoDAB module results in a more significant performance drop (Dice: 85.15%), reinforcing the value of language-driven guidance.

Architectural Component Evaluation: Substituting the cross-attention with point-wise addition reduces segmentation accuracy to 92.11%, and replacing the SSMix with a linear projection yields 91.72%. These results highlight the importance of structured attention and dynamic sequence modeling in multimodal integration.

7. Conclusion

In this research, we proposed a novel uncertainty-aware vision-language segmentation model designed to enhance medical image segmentation. Our model integrates visual and textual data through advanced cross-modal learning techniques, utilizing proposed key modules such as the Modality Decoding Attention Block (MoDAB) and the State Space Mixer (SSMix). These modules significantly improve segmentation accuracy by capturing both spatial and semantic information. Additionally, we proposed the Spectral-Entropic Uncertainty (SEU) Loss function, which guides the model to account for uncertainty during training, enhancing spatial precision and domain-specific visual-linguistic alignment. Comprehensive experiments on multiple medical datasets demonstrated that our model, equipped with the SEU loss, outperforms existing state-of-the-art methods in both accuracy and computational efficiency. These results underscore the potential of our approach to advance medical image analysis, offering more reliable and interpretable segmentation for clinical decision-making.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241. 1, 2, 7

[2] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018. 1, 2, 7

[3] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*. Springer, 2018, pp. 3–11. 1, 2, 7

[4] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218. 1, 2, 7

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763. 1, 2, 7

[6] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International conference on machine learning*. PMLR, 2021, pp. 5583–5594. 1, 2, 7

[7] Y. Zhong, M. Xu, K. Liang, K. Chen, and M. 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*. Springer, 2023, pp. 724–733. 1, 2, 7

[8] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023. 1, 2

[9] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021. 2, 7

[10] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, and Y. Zhou, "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841524002056 2, 7

[11] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, pp. 2441–2449. 2, 7

[12] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2401.04722 2, 7

[13] J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, C. Li, Y. Liang, G. Shi, Y. Yu, S. Zhang, H. Zheng, and S. Wang, " Swin-UMamba: Mamba-based UNet with ImageNet-based pretraining ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. LNCS 15009. Springer Nature Switzerland, October 2024. 2, 7

[14] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://arxiv.org/abs/2111.00396 2

[15] M. Poli, S. Massaroli, E. Nguyen, D. Y. Fu, T. Dao, S. Baccus, Y. Bengio, S. Ermon, and C. Ré, "Hyena hierarchy: Towards larger convolutional language models," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: https://arxiv.org/abs/2302.10866 2

[16] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," in *International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: https://arxiv.org/abs/2307.08691 2

[17] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine learning for healthcare conference*. PMLR, 2022, pp. 2–25. 2, 7

[18] S.-C. Huang, L. Shen, M. P. Lungren, and S. 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*, 2021, pp. 3942–3951. 2, 7

[19] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong *et al.*, "Large-scale domain-specific pretraining for biomedical vision-language processing," *arXiv preprint arXiv:2303.00915*, vol. 2, no. 3, p. 6, 2023. 2, 7

[20] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. Torr, "Lavt: Language-aware vision transformer for referring image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 155–18 165. 2, 7

[21] Z. Li, Y. Li, Q. Li, P. Wang, D. Guo, L. Lu, D. Jin, Y. Zhang, and Q. Hong, "Lvit: Language meets vision transformer in medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 96–107, 2024. 2, 6, 7

[22] M. Xu, T. Xiao, Y. Liu, H. Tang, Y. Hu, and L. Nie, "Cmirnet: Cross-modal interactive reasoning network for referring image segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[23] S. Ouyang, H. Wang, S. Xie, Z. Niu, R. Tong, Y.-W. Chen, and L. Lin, "Slvit: Scale-wise language-guided vision transformer for referring image segmentation." in *IJCAI*, 2023, pp. 1294–1302. 2, 7

[24] M. Li, M. Meng, S. Ye, M. Fulham, L. Bi, and J. Kim, "Language-guided medical image segmentation with target-informed multi-level contrastive alignments," *arXiv preprint arXiv:2412.13533*, 2024. 2

[25] X. Huang, H. Li, M. Cao, L. Chen, C. You, and D. An, "Cross-modal conditioned reconstruction for language-guided medical image segmentation," *IEEE Transactions on Medical Imaging*, 2024. 2, 7

[26] C. Li, H. Zhu, R. I. Sultan, H. B. Ebadian, P. Khanduri, C. Indrin, K. Thind, and D. Zhu, "Mulmodseg: Enhancing unpaired multi-modal medical image segmentation with modality-conditioned text embedding and alternating training," *arXiv preprint arXiv:2411.15576*, 2024. 2

[27] Y. Hu, Q. Wang, W. Shao, E. Xie, Z. Li, J. Han, and P. Luo, "Beyond one-to-one: Rethinking the referring image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4067–4077. 2, 7

[28] J. Wu, X. Li, X. Li, H. Ding, Y. Tong, and D. Tao, "Toward robust referring image segmentation," *IEEE Transactions on Image Processing*, vol. 33, pp. 1782–1794, 2024. 2, 7

[29] N. K. Tomar, D. Jha, U. Bagci, and S. Ali, "Tganet: Text-guided attention for improved polyp segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 151–160. 2, 7

[30] J. Hu, Y. Li, H. Sun, Y. Song, C. Zhang, L. Lin, and Y.-W. Chen, "Lga: A language guide adapter for advancing the sam model's capabilities in medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 610–620. 2, 7

[31] G. Chen, "Text-guided multi-stage cross-perception network for medical image segmentation," *arXiv preprint arXiv:2506.07475*, 2025. 2

[32] X. Zhang, B. Ni, Y. Yang, and L. Zhang, " MAdapter: A Better Interaction between Image and Language for Medical Image Segmentation ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. LNCS 15009. Springer Nature Switzerland, October 2024. 2, 7

[33] T. Zeevi, L. H. Staib, and J. A. Onofrey, "Enhancing uncertainty estimation in semantic segmentation via monte-carlo frequency dropout," *arXiv preprint arXiv:2501.11258*, 2025. 2

[34] M. Antico, G. Bruno, E. Faggiano *et al.*, "Evaluating uncertainty quantification in medical image segmentation: A multi-dataset, multi-algorithm study," *Applied Sciences*, vol. 14, no. 21, p. 10020, 2022. 2

[35] S. Sedai, D. Mahapatra, and R. Garnavi, "Uncertainty guided semi-supervised segmentation of retinal layers in oct images," *Medical Image Analysis*, vol. 57, pp. 226–236, 2019. 3

[36] A. G. Roy, N. Navab, and C. Wachinger, "Bayesian quicknat: Model uncertainty in deep whole-brain segmentation for structure-wise quality control," in *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2019, pp. 653–661. 3

[37] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 11 976–11 986. 3

[38] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay, "Making the most of text semantics to improve biomedical vision–language processing," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 1–21. 3

[39] K. Poudel, M. Dhakal, P. Bhandari, R. Adhikari, S. Thapaliya, and B. Khanal, "Exploring transfer learning in medical image segmentation using vision-language models," *arXiv preprint arXiv:2308.07706*, 2023. 6

[40] A. Degerli, S. Kiranyaz, M. E. Chowdhury, and M. Gabbouj, "Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images," *arXiv preprint arXiv:2202.10185*, 2022. 6

[41] S. P. Morozov *et al.*, "Mosmeddata: Chest ct scans with covid-19 related findings dataset," *arXiv preprint arXiv:2005.06465*, 2020. 6

[42] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, no. 1, pp. 1–13, 2020. 6

[43] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462. 6

[44] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017. [Online]. Available: http://arxiv.org/abs/1711.05101 7

[45] ——, "SGDR: stochastic gradient descent with restarts," *CoRR*, vol. abs/1608.03983, 2016. [Online]. Available: http://arxiv.org/abs/1608.03983 7