HiMix : Hierarchical Visual-Textual Mixing Network for Lesion Segmentation

Soojin Hwang* Jaeyoon Sim* Won Hwa Kim
Pohang University of Science and Technology (POSTECH)
{soojin0622, simjy98, wonhwa}@postech.ac.kr

Abstract

Lesion segmentation is an essential task in medical imaging to support diagnosis and assessment of pathologies. While deep learning models have shown success in various domains, their reliance on large-scale annotated datasets limits its applicability in the medical domain due to labeling cost. To address this issue, recent studies in medical image segmentation have utilized clinical texts as complementary semantic cues without additional annotations. However, most existing methods utilize a single textual embedding and fail to capture hierarchical interactions between language and visual features, which limits their ability to leverage fine-grained cues essential for precise and detailed segmentation. In this regime, we propose Hierarchical Visual-Textual Mixing Network (HiMix), a novel multi-modal segmentation framework that mixes multi-scale image and text representations throughout the mask decoding process. HiMix progressively injects hierarchical text embedding, from high-level semantics to fine-grained spatial details, into corresponding image decoder layers to bridge the modality gap and enhance visual feature refinement at multiple levels of abstraction. Experiments on the QaTa-COV19, MosMedData+ and Kvasir-SEG datasets demonstrate that HiMix consistently outperforms uni-modal and multi-modal methods. Furthermore, HiMix exhibits strong generalization to unstructured textual formats, highlighting its practical applicability in real-world clinical scenarios.

1. Introduction

Medical semantic segmentation is crucial for precise identification and diagnosis of various pathological conditions. Recent deep learning-based automated segmentation methods [5–7, 13, 14, 19, 33, 37, 44, 46, 56] have achieved considerable success in assisting clinicians with lesion detection and treatment planning. However, these deep learning models rely heavily on large-scale annotated image datasets, which require substantial cost, time and effort to acquire. This limitation is particularly critical in the medical domain, where strict privacy constraints and necessity of expert annotations severely restrict the accessibility of large-scale data.

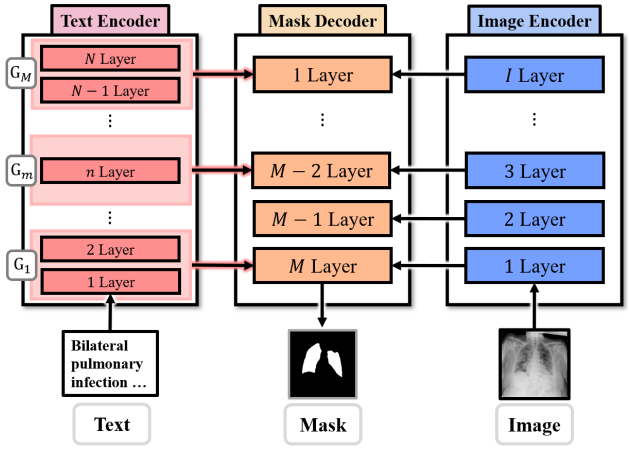


Figure 1. Illustration of text feature utilization strategy in our proposed approach. Multi-level linguistic features extracted from the text encoder are hierarchically aligned and passed to the corresponding decoder layers.

To improve the segmentation result with limited sample size, recent studies [4, 27, 55] have increasingly explored medical reports which are naturally paired with medical images in clinical archives. Medical reports often provide important clinical details, e.g., the exact number and location of infected areas, which are routinely captured in medical reports by radiologists with specialized terminology, e.g., "unilateral" or "bilateral" [24] but overlooked in visual information. Such an approach not only reduces the need for newly labeled data but also provides a semantic guide that helps resolve ambiguities in image interpretation.

Although these studies demonstrate the benefit of multi-modal integration and propose various fusion strategies, they mainly focus on selecting fusion stages and mechanisms, rather than exploring optimal linguistic representations for guiding segmentation. In doing so, they adopt a pre-trained language model, e.g., BERT [9], to obtain an

*Equal contribution.

abstract embedding in the final layer. As a result, they leave out the rich and detailed linguistic features embedded in other layers that are critical for spatially fine-grained segmentation.

This limitation is further emphasized by BERTology studies [20, 28, 36, 42], which reveal that transformer-based language models capture distinct types of information at different layers: lower layers capture surface patterns and local phrasing, middle layers capture syntactic dependencies, and upper layers represent abstract semantics. Relying solely on final-layer semantics is insufficient for spatially precise tasks such as medical segmentation, which require structured information distributed throughout the report. Such cues are embedded not only in the overall context but also in word choices, phrase structures, and grammatical patterns. For example, the report "Bilateral pulmonary infection, two infected areas, all left lung and upper middle lower right lung" conveys spatial context through terms such as "bilateral", count phrases such as "two infected areas", and structured expressions describing specific anatomical regions. Accurate interpretation of such information requires attending to multiple linguistic levels beyond the capacity of a single embedding layer.

The hierarchical nature of text embeddings encourages to reconsider how text information is leveraged in multimodal medical segmentation. Exploiting multi-level linguistic cues naturally aligns with the structure of standard image-based encoder-decoder architecture, which refines segmentation mask progressively from coarse semantic features at early stages to finer details later. Inspired by this hierarchical visual refinement, we argue that a similar multi-level utilization strategy should be applied to the text modality. Specifically, abstract semantics from higher text layers can guide coarse segmentation estimates in the early decoder stages. In contrast, detailed cues from lower and middle layers such as explicit counts, spatial relations, and anatomical references (e.g., "two infected areas" and "left lung") are crucial for precise boundary delineation in later stages. Aligning linguistic features with corresponding level of visual refinement enables these layered textual cues to meaningfully contribute to accurate and anatomically faithful segmentation.

In this regards, we propose **Hi**erarchical Visual-Textual **Mix**ing Network (HiMix), a novel multi-modal framework for lesion segmentation. HiMix effectively aligns and leverages hierarchical representations from both image and text modalities. For this, we design Dynamic Layer Fusion Module (DLFM), which aligns hierarchical textual representations with image features throughout the segmentation mask decoding stages. In parallel, we propose the Adaptive Spectrum Refinement Module (ASRM), which dynamically adjusts kernel bandwidths across decoding layers to capture global-to-local visual contexts more effectively. By harmonizing these refined visual features with hierarchical text representations, HiMix achieves fine-grained and semantically guided segmentation. Our approach differs from prior methods [4, 27, 55] that typically inject a single final-layer text embedding throughout the decoder. Instead, as shown in Fig. 1, HiMix partitions the text encoder into semantic levels (e.g., low-, mid-, and high-level) and aligns each with corresponding decoder stages. This design enables structured semantic alignment between text and image representation across decoding, leading to finer and more contextually grounded segmentation.

In summary, HiMix offers the following **key contributions: 1)** HiMix introduces a hierarchical fusion strategy that effectively aligns multi-level linguistic and visual representations for medical image segmentation. **2)** HiMix shows superior performance on lesion segmentation compared to the state-of-the-art methods. **3)** HiMix demonstrates adaptability through extensive experiment with other text formats. Extensive validation was performed on two independent lung infection benchmarks, QaTa-COV19 and MosMedData+ datasets, to assess the effectiveness and generalizability of our model.

2. Related Work

2.1. Medical Semantic Segmentation

Deep learning has driven significant progress in medical image segmentation, with numerous methods proposed over the years. Early foundational work predominantly relied on Convolutional Neural Networks (CNNs), demonstrating remarkable capabilities in capturing hierarchical visual features. The U-Net architecture [37] and its subsequent variations such as U-Net++ [56], Attention U-Net [33], and more advanced design such as nnU-Net [19], became standard benchmarks and have showcased the power of CNNs for local feature extraction and spatial reasoning in medical tasks.

To enhance segmentation performance, Transformer-based architectures have been introduced to capture global dependencies and long-range spatial interactions. Representative approaches such as TransUNet [6], UNETR [14], Swin-UNet [5], and Swin-UNETR [13] leverage self-attention mechanisms. UCTransNet [46] combines CNNs and Transformers to better balance local and global feature extraction. Despite these advances, most existing methods remain image-centric, relying solely on visual input without semantic cues from clinical knowledge, which hinders their capacity to interpret subtle visual patterns requiring contextual understanding. Moreover, high performance with image-only models often depends on large-scale and expert-annotated datasets, which are costly and difficult to obtain in medical domains. These challenges highlight the need for multi-modal approaches that combine visual data with clinically meaningful information, such as radiology reports.

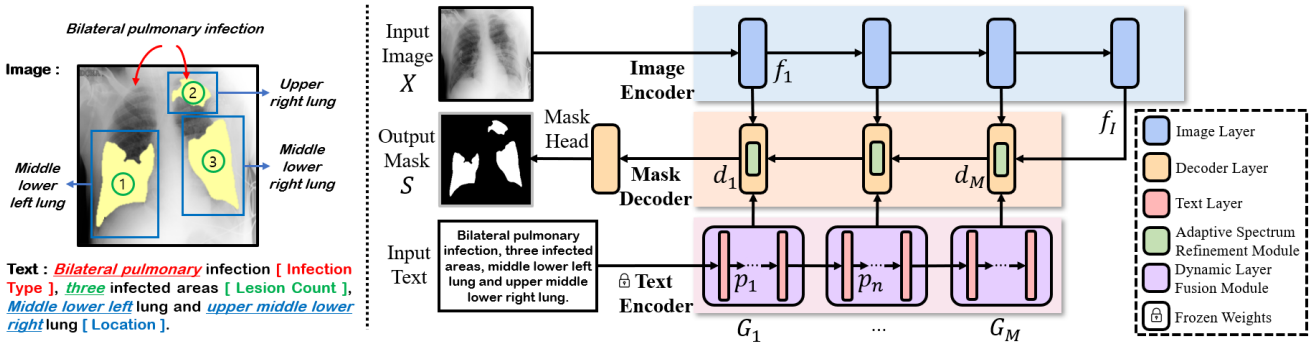


Figure 2. **Architecture of HiMix.** Left: Illustration of an image-prompt describing structured radiologic findings. Right: An input image X and text prompt are fed into the image and text encoder, respectively. The text feature p is refined through DLFM, while the image feature f is directly fed to the decoder, allowing visual features to be sequentially refined through ASRM to predict the output mask S.

2.2. Language-guided segmentation

Recently, multi-modal approaches have incorporated medical images with corresponding radiology reports [2, 4, 10, 25, 27, 32, 43, 55] to overcome the inherent limitations of image-only segmentation. These approaches utilize diagnostic cues in clinical text such as lesion counts, anatomical locations, and characteristics to enhance the interpretation of visual data. Specifically, TGANet [43] and AKGNet [10] explicitly mine attribute-level cues (e.g., counts) and inject them via auxiliary heads or cross-attention, while others fuse continuous text representations with visual features without explicit attribute mining. Incorporating such semantic information provides complementary guidance that improves the precision of image-based segmentation.

Focusing on the fusion direction, recent efforts have focused on developing effective fusion strategies that combine textual semantics with visual features. For instance, LViT [27] injects textual embeddings early in the visual processing pipeline to guide attention, GuideDecoder [55] conditions the image decoder on text features, to influence upsampling and segmentation. MMI-UNet [4] performs feature matching between modalities at intermediate layers. However, these methods overlook the hierarchical structure of linguistic cues across multiple layers, which limits fine-grained semantic signals such as anatomical relations and spatial structure. Unlike prior work, HiMix aligns multi-level linguistic and visual cues in a hierarchical manner to achieve refined and precise segmentation.

3. Method

3.1. Overview of HiMix

Given an image-text pair, i.e., chest CT scan and medical report, our goal is to generate a context-aware pixel-wise mask that localizes disease-specific regions. To achieve this, we design HiMix as a comprehensive multi-modal segmentation framework consisting of three core components including 1) an image encoder, 2) a text encoder, and 3) a mask decoder, as illustrated in Fig. 2.

The image encoder first extracts multi-scale visual features from an image X with a CNN backbone. Simultaneously, the text encoder encodes a medical report through a pre-trained language model to yield layer-wise textual embeddings P, which are grouped and selectively aggregated by DLFM to align hierarchical textual semantics with corresponding decoder stages. Within the decoder, the ASRM dynamically enhances visual features from global to local contexts based on their spatial frequency components. The decoder then fuses these refined visual and textual representations using attention-based interactions to generate the final segmentation mask S.

3.2. Image & Text Encoder

Given an image X, we extract multi-level image features $F=\{f_i\}_{i=1}^{I}$ where $f_i \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i}$ using ConvNeXt-Tiny [29] with I layers as an image encoder. Here, H and W denote the height and width of X, and C_i is the feature dimensions at the i-th layer. For an input text, we adopt the pre-trained text encoder (e.g., CXR-BERT [3]) to extract a sequence of hidden features $P=\{p_n\}_{n=1}^{N}$ from N layers with the text embedding $p_i \in \mathbb{R}^{L \times C_P}$ with length of L in dimension of C_P. These features encode different levels of semantic information from the input text, capturing both local and global context, with higher indices representing more abstract and high-level information.

Dynamic Layer Fusion. To better align with the structure of a decoder and exploit the hierarchical text embeddings, we introduce a Dynamic Layer Fusion Module (DLFM) which constructs the representative textual embedding of each group based on semantic levels. Given a mask decoder with M layers, we evenly split the hidden features P into M groups, denoted as $G_m = \{p_j | j \in J_m\}$ for $1 \leq m \leq M$, where $J_m = \{j | j \in [\lceil \frac{N}{M} \cdot (m-1) + 1 \rceil, \lceil \frac{N}{M} \cdot m \rceil]\}$.

Then, G_M is fed into the first decoder layer, while G_{M-1} to G_1 are sequentially passed to the following layers. For

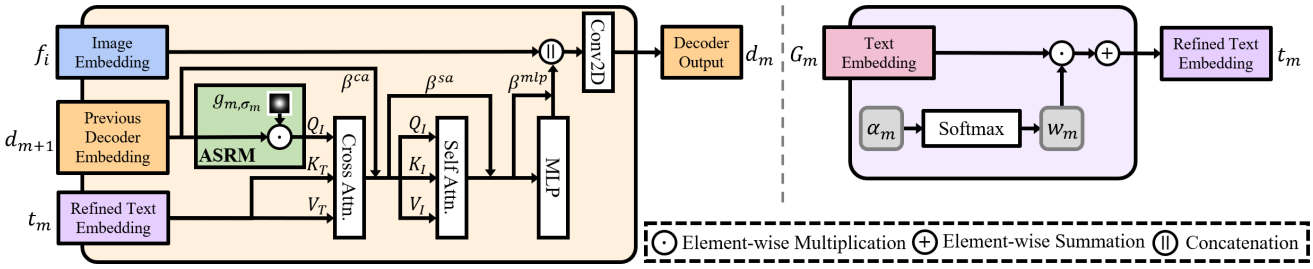


Figure 3. **Illustration of key modules in HiMix.** Left: Decoder with adaptive spectrum refinement module (ASRM). Right: Dynamic layer fusion module (DLFM).

each G_m, HiMix dynamically learns the contribution w_j of the j-th hierarchical text embedding within the m-th group, parameterized by $\boldsymbol{\alpha}_m$ as $w_j = \frac{e^{\alpha_j}}{\sum_{j \in J_m} e^{\alpha_j}}$. Here, we denote the learnable logit vector for G_m by $\boldsymbol{\alpha}_m = \{\alpha_j\}_{j \in J_m}$, produced by a lightweight gating head.

Finally, the refined text embedding $t_m \in \mathbb{R}^{L \times C_P}$ at each decoder stage is computed as

$$t_m = \sum_{j \in J_m} w_j \cdot p_j. \tag{1}$$

The text embedding for each group is then passed to the corresponding decoder layer, ensuring that the text features are aligned with the visual context at each decoding stage.

3.3. Segmentation Mask Decoder

To begin the decoding process, the initial input d_{M+1} of the top-most decoder is set to the global image feature f_I. Given the i-th image embedding f_i, the m-th refined text embedding t_m, and the $(m+1)$-th decoder output d_{m+1}, the decoder generates the m-th output feature $d_m \in \mathbb{R}^{(H \times W) \times C_m}$, which integrates both image and text characteristics, as shown in Fig. 3. Before the integration, text and image token dimensions are aligned by projecting t_m to $\hat{t}_m$ using a transformation defined as

$$\hat{t}_m = \mathrm{ReLU}(\mathrm{Conv}(t_m W_t)) \in \mathbb{R}^{L' \times C_i}. \tag{2}$$

Here, L' represents the number of tokens after projection, W_t is a learnable parameter, and $\mathrm{Conv}(\cdot)$ denotes a 1×1 convolutional layer. The obtained $\hat{t}_m$ will be matched with the visual feature later.

Adaptive Spectrum Refinement. Several studies on neural networks in the spectral domain [39, 40, 50, 52] suggest that low-frequency components convey high-level semantic features, whereas fine details in the high-frequency components are often overlooked. Motivated by these insights, we propose Adaptive Spectrum Refinement Module (ASRM) that enhances coarse semantics or fine-grained details adaptively at each decoder stage, promoting balanced multi-level representations. The refined visual feature $\hat{d}_m$ of the m-th decoder layer is obtained from d_m as

$$\hat{d}_m = \mathrm{IFFT}(\mathrm{FFT}(d_m) \odot g_{m,\sigma_m}) + d_m \tag{3}$$

where $\odot$ is element-wise multiplication, g_{m,σ_m} is a Gaussian filtering coefficient map with kernel bandwidth σ_m, matching d_m's spatial dimensions and FFT and IFFT represent Fourier Transform and Inverse Fourier Transform, respectively. Here, σ_m is learnable, flexibly capturing of local-to-global visual properties.

With refined visual feature $\hat{d}_m$ and textual embedding $\hat{t}_m$, the multi-head cross-attention (MHCA) mechanism [45] fuses textual guidance with visual information, producing a text-guided visual representation as

$$d_m^{ca} = \hat{d}_m + \beta^{ca}(\mathrm{LN}(\mathrm{MHCA}(\hat{d}_m, \hat{t}_m)) \tag{4}$$

where $\mathrm{LN}(\cdot)$ is layer normalization [1], and β^{ca} is a trainable parameter controlling the weight of the residual connection. Next, the obtained representation d_m^{ca} is inputted to a multi-head self-attention (MHSA) operation to enhance contextual relations within the decoder embedding itself as

$$d_m^{sa} = d_m^{ca} + \beta^{sa}(\mathrm{LN}(\mathrm{MHSA}(d_m^{ca})) \tag{5}$$

where β^{sa} is also a learnable parameter. Then, the enhanced visual representation d_m^{sa} is then passed to a Multi-Layer Perceptron (MLP) with an activation and residual connection, and produce a final visual feature as

$$d_m = d_m^{sa} + \beta^{mlp}(\mathrm{MLP}(d_m^{sa})) \tag{6}$$

where β^{mlp} is a learnable weight that adjusts the influence of the residual connection. Afterwards, the resulting mixed multi-modal feature d_m is reshaped and upsampled to derive d_m^u, which is denoted as

$$d_m^u = \mathrm{Upsample}(\mathrm{Reshape}(d_m)) \in \mathbb{R}^{H' \times W' \times C_m} \tag{7}$$

where H' and W' denote the upsampled height and width, respectively. To compute the final decode output d_m^o, the d_m^u is concatenated with $f_m \in \mathbb{R}^{H' \times W' \times C_m}$ on the channel dimension, where f_m is the low-level features obtained from the visual encoder via skip connection, derived as

$$d_m^o = \mathrm{ReLU}(\mathrm{Conv}([d_m^u, f_m]) \in \mathbb{R}^{H' \times W' \times C_{m-1}} \tag{8}$$

where $[\cdot, \cdot]$ represents channel-wise concatenation.

Table 1. Quantitative comparison on segmentation of uni-modal (top) and multi-modal (middle) learning baselines, and HiMix (bottom). The best and second-best results are highlighted in **bold** and underlined, respectively.

Approach	Type	Method	Param ↓ (M)	QaTa-COV19		MosMedData+		Kvasir-SEG	
				DSC ↑	IoU ↑	DSC ↑	IoU ↑	DSC ↑	IoU ↑
Uni-Modal (Image-Only)	CNN	UNet [37]	14.8	79.02	69.46	64.60	50.73	82.94	74.47
		UNet++ [56]	74.5	79.62	70.25	71.75	58.39	80.43	72.13
		AttnUNet [33]	34.9	79.31	70.04	66.34	52.82	81.31	73.74
		nnUNet [19]	19.1	80.42	70.81	72.59	60.36	85.06	74.01
	Transformer	Swin-UNet [5]	82.3	78.07	68.34	63.29	50.19	75.97	67.45
	Hybrid	UCTransNet [46]	65.6	79.15	69.60	65.90	52.69	78.21	65.25
		TransUNet [6]	105.0	78.63	69.13	71.24	58.44	79.67	71.14
	Mamba	VMUNet [38]	31.0	86.31	75.92	75.85	61.09	79.54	66.04
		H-vmunet [49]	31.0	86.26	75.84	76.37	61.78	82.25	69.85
Multi-Modal (Image-Text)	CNN	GLoRIA [17]	45.6	79.94	70.68	72.42	60.18	84.73	73.51
		ConVIRT [53]	35.2	79.72	70.58	72.06	59.73	84.27	72.82
		RecLMIS [18]	23.7	85.22	77.00	77.48	65.07	87.73	78.15
	Transformer	CLIP [35]	87.0	79.81	70.66	71.97	59.64	81.27	68.45
		MedCLIP [47]	137.0	86.54	76.27	69.14	52.84	78.84	65.08
		DMMI [15]	131.5	79.25	68.87	72.57	60.78	74.79	59.73
		RefSegformer [48]	195.0	84.09	75.48	74.98	61.70	79.78	66.37
		MedSAM [30]	**4.5**	78.49	69.11	54.22	42.22	86.69	79.24
	Hybrid	ViLT [23]	87.4	79.63	70.12	72.36	60.15	70.33	54.23
		LAVT [51]	118.6	79.28	69.89	73.29	60.41	70.59	54.55
		LViT [27]	29.7	83.66	75.11	74.57	61.33	87.17	77.26
		SLViT [34]	114.6	84.13	75.66	75.01	61.83	86.61	76.38
		GuideDecoder [55]	44.0	89.78	81.45	77.75	63.60	88.31	79.07
		MMI-UNet [4]	56.2	90.88	83.28	78.42	64.50	91.43	84.57
		HiMix (Ours)	44.7	**91.17**	**83.78**	**79.44**	**65.90**	**92.18**	**85.50**

Finally, the final output S (i.e., mask) is subsequently derived from the final decoder feature map d_1^o as

$$S = \text{Sigmoid}(\text{Conv}(\text{Upsample}(d_1^o))), \tag{9}$$

which produces pixel-wise segmentation results. Overall, the proposed architecture promotes effective utilization of hierarchical textual and visual cues, yielding accurate and context-aware lesion segmentation.

4. Experiments

4.1. Datasets and Experimental Settings

Dataset. Three benchmarks, i.e., Qata-COV19 (QATA) [8], MosMedData+ (MOSMED) [31], and Kvasir-SEG (KVASIR) [21], were used in our experiments to evaluate the performance of HiMix. The QATA dataset includes 9258 COVID-19 chest X-ray images, which was partitioned into training, validation, and testing sets, with 5716, 1429, and 2113 samples, respectively. Similarly, the MOSMED dataset consists of 2729 CT scan slices of lung infections, partitioned into 2183 for training, 273 for validation, and 273 for testing. For the KVASIR dataset, they comprise 1000 colonoscopy frames with pixel-wise polyp masks, split into training, validation and testing with a ratio of 8:1:1. All datasets share similar text annotations, highlighting key clinical factors such as abnormality count, visual attributes, and their approximate location within an image.

Experimental Setup. Following the setup in [27] and [4], we designed medical image segmentation experiments to detect disease-related regions from image-text pairs. For quantitative evaluation, we employed the Dice coefficient (DSC) and IoU metrics to measure overlap between predicted and ground truth masks, as in [27]. For optimization, we employed a combination of Dice loss [41] and Cross-Entropy loss [54]. Additional details on the model implementation are provided in supplementary material.

4.2. Segmentation Result Comparisons

Quantitative Comparisons. Table 1 reports the results of HiMix against recent uni- and multi-modal baselines spanning diverse architectures (e.g., Transformers and state-space models). The top of Table 1 presents uni-modal baselines that rely solely on visual inputs [5, 6, 19, 33, 37, 38, 46, 49, 56], while the bottom contains multi-modal methods utilizing medical images with associated clinical text [4, 15, 17, 18, 23, 27, 30, 34, 35, 47, 48, 51, 53, 55]. HiMix surpassed all listed uni- and multi-modal baselines on all datasets and achieves state-of-the-art performance in both DSC and IoU metrics, highlighting the overall superiority of HiMix across medical segmentation benchmarks.

While uni-modal baselines exhibit reasonable segmentation performance, their DSC and IoU scores remain noticeably lower than those of HiMix. Specifically, HiMix achieves substantial DSC gains of 4.86%p on QATA, 3.07%p on MOSMED and 9.93%p on KVASIR over the best-performing uni-modal method. This performance margin clearly highlights the significant advantage of incorporating external clinical information such as lesion character-

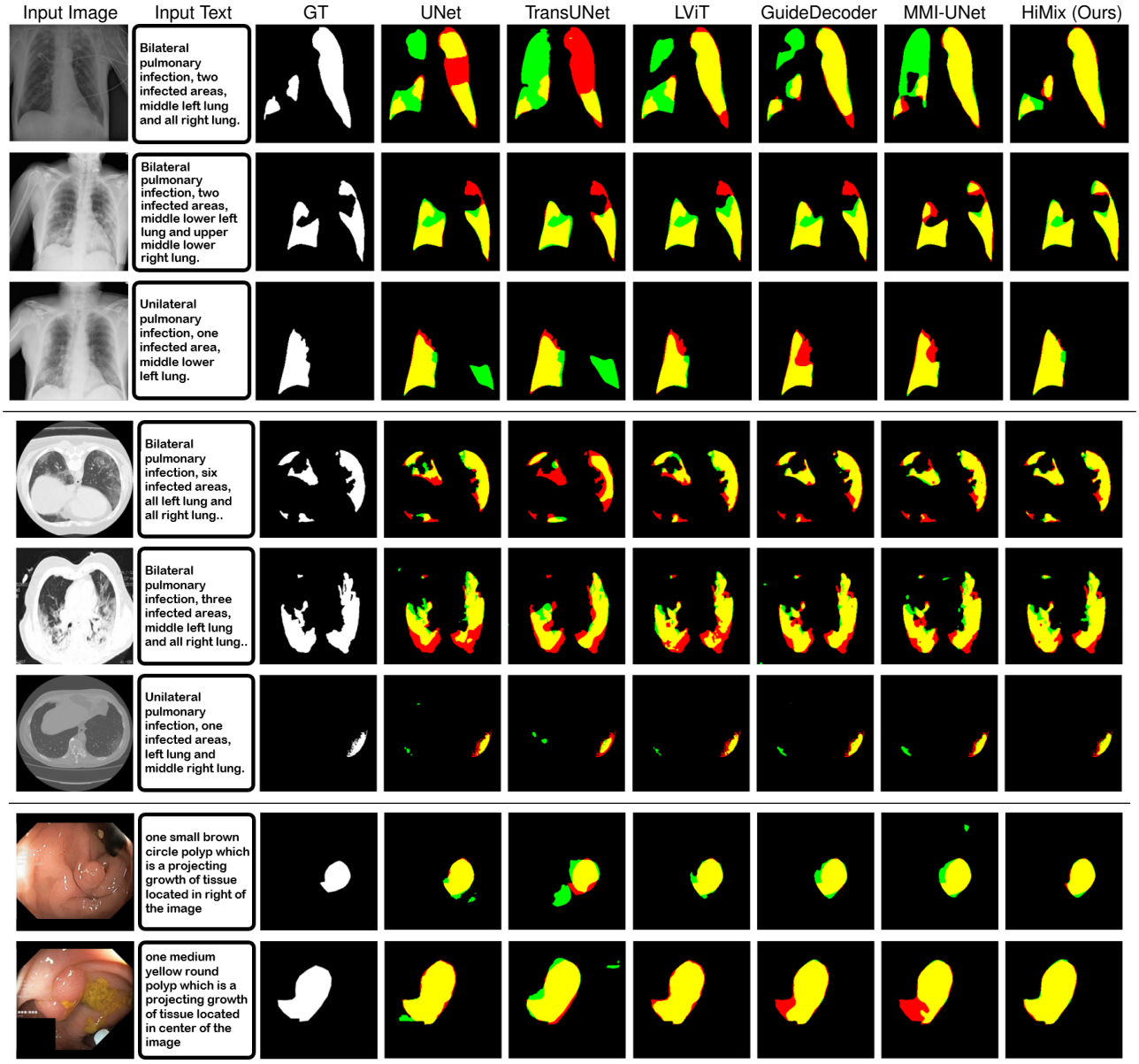


Figure 4. **Visualization of segmentation results.** Results on the QaTa-COV19 (top), MosMedData+ (middle) and Kvasir-SEG (bottom) datasets. Yellow, red, and green represent true positive, false negative, and false positive, respectively.

istics, which is entirely absent in uni-modal settings.

In general, multi-modal approaches consistently outperform uni-modal baselines, highlighting the utility of clinical text. Among multi-modal methods, HiMix achieves a DSC of 91.17% on QATA, 79.44% on MOSMED, and 92.18%p on KVASIR and outperforms recent SOTA multi-modal models such as MMI-UNet [4] by margins of 0.29%p on QATA, 1.02%p on MOSMED and 0.75%p on KVASIR, respectively. These improvements demonstrate the effectiveness of our dual-module design in fully leveraging multi-level textual and visual information. Moreover, HiMix achieves superior performance with only 44.7M parameters, significantly fewer than Transformer-based models, and even fewer than the second-best model [4] with 56.2M.

Qualitative Comparisons. Fig. 4 shows qualitative comparisons of HiMix against five baselines (i.e., two uni-modal and three multi-modal methods) on QATA and MOSMED. Here, yellow regions indicate true positives where the predicted segmentation overlaps with the ground truth, red regions denote false negatives corresponding to missed lesion areas, and green regions represent false positives where predictions exceed the true boundaries. These results intuitively assess segmentation accuracy, which we further analyze through four representative cases below.

Table 2. Ablation study on the modules of HiMix. The highest performance is achieved when both ASRM and DLFM are included.

Module		QATA		MOSMED		KVASIR	
ASRM	DLFM	DSC ↑	IoU ↑	DSC ↑	IoU ↑	DSC ↑	IoU ↑
×	×	91.06	83.58	78.28	64.31	91.31	83.34
✓	×	91.10	83.66	78.90	65.16	91.53	84.70
×	✓	91.13	83.71	78.70	64.88	92.10	85.05
✓	✓	**91.17**	**83.78**	**79.44**	**65.90**	**92.18**	**85.50**

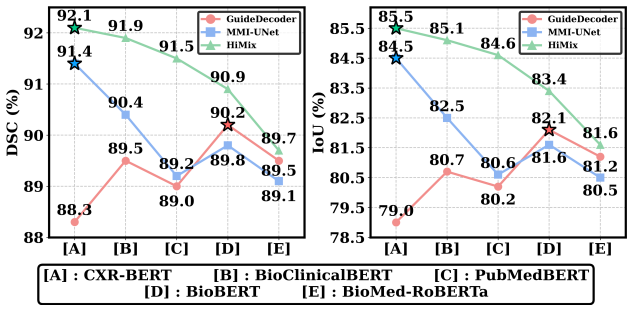


Figure 5. Ablation study of text encoders on the KVASIR dataset. The best result for each method is marked with a star, and across all text encoders, HiMix consistently achieves the best performance.

Case 1) Lesion size variation. Across all datasets, HiMix accurately segments both large and small lesions with high structural precision and minimal noise. It preserves shape fidelity in large lesions and maintains accurate boundaries even for small or faint ones. In contrast, baseline methods often miss small regions or introduce spurious noise, resulting in incomplete or distorted segmentation.

Case 2) Lesion count variation. In the second and last rows of the MOSMED results, where reports describe spatial constraints such as "three infected areas" and "unilateral" involvement, HiMix correctly captures the described lesion counts and distributions and reflects them, unlike baselines which often merge or misplace regions.

Case 3) Anatomical distribution. For example, in the first row of both QATA and KVASIR, where the reports specify detailed anatomical locations such as "middle left lung" or "small, right", HiMix follows these spatial constraints by segmenting only the indicated lobes. In contrast, other methods often miss relevant regions or over-extend beyond the reported boundaries. This is enabled by HiMix's alignment of hierarchical linguistic cues with visual decoding stages, allowing more faithful reflection of count and location constraints in the report.

Case 4) Suppression of false negatives. Specifically, in the second and last rows for QATA and the last row for KVASIR examples, while HiMix detects subtle regions with high sensitivity, other methods often fail to identify these infected areas, leading to partial or missing predictions. These results demonstrate that HiMix captures even weak signals of infection and ensures inclusion of all infected areas, which is important in clinical settings.

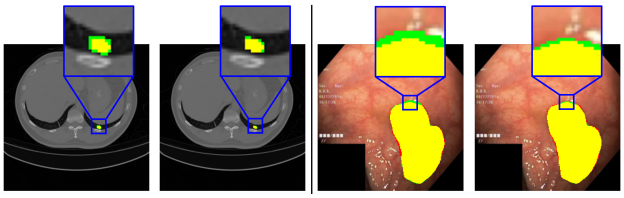
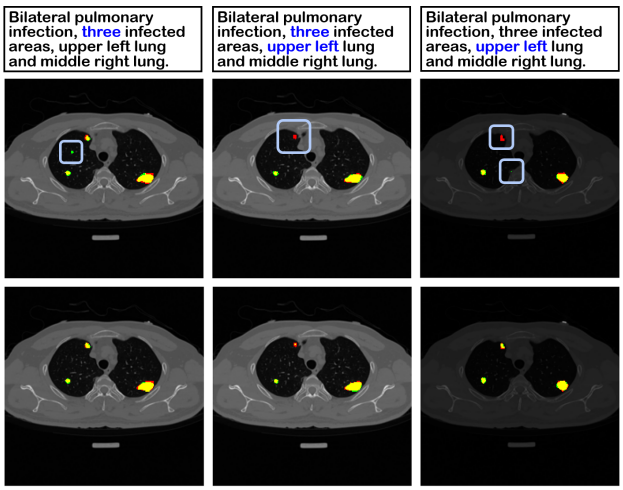


Figure 6. Qualitative comparison on the MOSMED (left) and KVASIR (right), showing results without ASRM (left) and with ASRM (right) for each dataset. HiMix with ASRM reduces over-segmentation and yields more precise boundaries.



(a) (b) (c)

Figure 7. Ablation study on the importance of anatomical relationships in text representation on the MOSMED dataset. A comparison of segmentation predictions from a model that uses only the final layer of the text encoder (top) and HiMix (bottom).

4.3. Analyses and Discussions

Ablation Study on Key Modules. To evaluate the effect of each module, we conducted an ablation study, which is reported in Table 2. Specifically, on the MOSMED dataset, removing DLFM led to a performance drop of 0.54%p in DSC and 0.74%p in IoU, while removing ASRM resulted in 0.64%p decrease in DSC and 1.02%p in IoU, respectively. These results demonstrate that both modules independently improve segmentation accuracy. Notably, although the absence of both modules drops DSC to 91.06% on QATA, 78.28% on MOSMED and 91.31% on KVASIR, our approach still outperforms the majority of baselines in the experiments. This suggests that the overall design of HiMix, particularly leveraging full-layer language features, remains effective even without the two refinement modules.

Ablation Study on Text Encoders. To assess the benefit of the proposed hierarchical text–image architecture, Fig. 5 presents an ablation of pretrained text encoders for HiMix. We adopt CXR-BERT [3] as the default text encoder and ad-

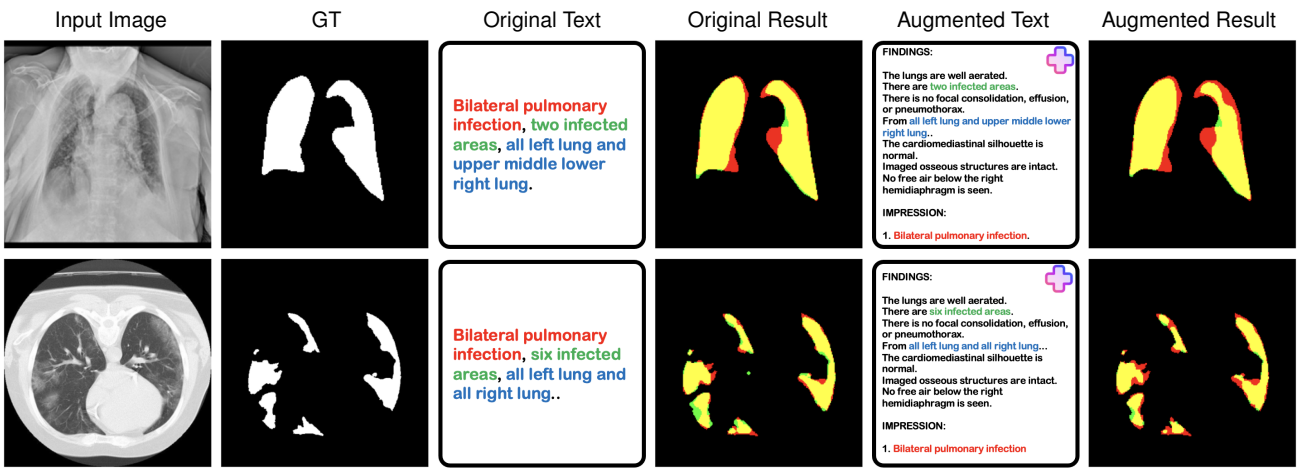


Figure 8. **Effect of text augmentation for segmentation.** The original text is structured, while the augmented text mimics medical reports from MIMIC-CXR. HiMix consistently captures lesion regions even with augmented text.

ditionally evaluate BioClinicalBERT [16], PubMedBERT [11], BioBERT [26], and BioMed-RoBERTa [12]. Across all encoders, HiMix consistently surpasses the recent SOTA method such as MMI-UNet [4] and GuideDecoder [55] on both DSC and IoU. These results indicate that the proposed hierarchical text–image fusion leverages multi-level linguistic cues rather than relying on a single final-layer embedding, yielding robust gains that are largely insensitive to the specific choice of text encoders.

Analysis of Spectrum Refinement. Fig. 6 shows qualitative segmentation results on the MOSMED and KVASIR datasets to highlight the effect of the proposed ASRM. Without ASRM (i.e., left column), the models tend to produce over-segmentations, leading to excessive false positives (green) and imprecise boundaries. In contrast, when ASRM is incorporated (i.e., right column), the predictions become more compact and better aligned with the ground truth, substantially reducing over-segmentation and improving boundary accuracy. This improvement across diverse scenarios demonstrates the generalizability of ASRM to heterogeneous medical datasets. Overall, HiMix suppress spurious activations while maintaining high sensitivity, yielding more reliable segmentations.

Analysis on Anatomical Relationships. Fig. 7 presents a qualitative ablation study on the capability of capturing anatomical relationships of HiMix. For this, we compare HiMix with a variant that uses only a final-layer text embedding. While the single-embedding variant incorrectly predicts four lesions instead of three in (a), our model precisely localizes all three areas. Similarly, in (b), our model accurately recovers all three lesions in the upper left lung by leveraging this key anatomical cue (i.e, upper left) from the text, unlike the ablated model which only detects two regions. Finally, in (c), the ablated model fails to utilize the text information, leading to spatial drift, whereas HiMix localizes the lesion by reflecting the spatial clue 'upper left'. Overall, these results validate that DLFM effectively captures complex anatomical relationships between text and images, which is critical for accurate clinical segmentation.

Generalization of HiMix. Unlike real medical reports, the text annotations are structured prompts, and thus their clinical applicability in real-scenarios can be questioned. To verify this, as shown in Fig. 8, we created an augmented text test data to resemble real medical report format in MIMIC-CXR free-text radiology reports [22] with varying length and structure while preserving core contents. Despite being trained on structured text, HiMix showed comparable results when tested with the augmented text, demonstrating robustness of HiMix and applicability beyond rigid formats. For example, in the second row, while the result is less detailed than with the original text, it accurately segments six infected areas instead of seven. This suggests that more complex text improves global information, such as the number of infected areas, but trades off precision around boundaries due to increased contextual diversity.

5. Conclusion

In this work, we proposed HiMix, a novel multi-modal segmentation framework that effectively aligns and leverages hierarchical representations from both image and text modalities. HiMix dynamically extracts and refines essential information from both modalities to ensure a hierarchical integration of high-level semantics and fine-grained details. Experiments on diverse medical segmentation benchmarks demonstrate that HiMix consistently outperforms state-of-the-art models, which validates the advantage of hierarchical design. Moreover, HiMix further demonstrates strong adaptability to diverse text formats, showcasing its potential for practical applicability to real-scenarios.

Acknowledgements

This research was supported by RS-2019-II191906 (AI Graduate Program at POSTECH, 10%), IITP-2025-RS-2024-00437866 (40%), and RS-2022-II220290 (50%).

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Riddhish Bhalodia, Ali Hatamizadeh, Leo Tam, et al. Improving pneumonia localization via cross-attention on medical images and reports. In *Medical Image Computing and Computer Assisted Intervention*. Springer, 2021. 3

[3] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*. Springer, 2022. 3, 7

[4] 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*. Springer, 2024. 1, 2, 3, 5, 6, 8

[5] Hu Cao, Yueyue Wang, Joy Chen, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*. Springer, 2022. 1, 2, 5

[6] Jieneng Chen, Jieru Mei, Xianhang Li, et al. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 2024. 2, 5

[7] Hyuna Cho, Injun Choi, Suha Kwak, and Won Hwa Kim. Interactive network perturbation between teacher and students for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 1

[8] Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, et al. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *IEEE International Conference on Image Processing (ICIP)*, pages 2306–2310. IEEE, 2022. 5

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American chapter of the association for computational linguistics*, 2019. 1

[10] Qing En and Yuhong Guo. Akgnet: Attribute knowledge guided unsupervised lung-infected area segmentation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2024. 3

[11] Yu Gu, Robert Tinn, Hao Cheng, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 2021. 8

[12] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, 2020. 8

[13] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*. Springer, 2021. 1, 2

[14] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, et al. Unetr: Transformers for 3d medical image segmentation. In *IEEE/CVF winter conference on applications of computer vision*, 2022. 1, 2

[15] Yutao Hu, Qixiong Wang, Wenqi Shao, et al. Beyond one-to-one: Rethinking the referring image segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2023. 5

[16] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019. 8

[17] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, et al. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *IEEE/CVF international conference on computer vision*, 2021. 5

[18] Xiaoshuang Huang, Hongxiang Li, Meng Cao, et al. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024. 5

[19] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, et al. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 2021. 1, 2, 5

[20] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019. 2

[21] Debesh Jha, Pia H Smedsrud, Michael A Riegler, et al. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*. Springer, 2019. 5

[22] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 2019. 8

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*. PMLR, 2021. 5

[24] David B Larson, Alex J Towbin, Rebecca M Pryor, et al. Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. *Radiology*, 2013. 1

[25] Go-Eun Lee, Seon Ho Kim, Jungchan Cho, et al. Text-guided cross-position attention for segmentation: Case of medical image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023. 3

[26] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2020. 8

[27] Zihan Li, Yunxiang Li, Qingde Li, et al. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 2023. 1, 2, 3, 5

[28] Nelson F Liu, Matt Gardner, Yonatan Belinkov, et al. Linguistic knowledge and transferability of contextual representations. *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019. 2

[29] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, et al. A convnet for the 2020s. In *IEEE/CVF conference on computer vision and pattern recognition*, 2022. 3

[30] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 2024. 5

[31] Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, et al. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *medRxiv*, 2020. 5

[32] Philip Müller, Georgios Kaissis, Congyu Zou, et al. Radiological reports improve pre-training for localized imaging tasks on chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022. 3

[33] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, et al. Attention u-net: Learning where to look for the pancreas. *Medical Imaging with Deep Learning*, 2018. 1, 2, 5

[34] Shuyi Ouyang, Hongyi Wang, Shiao Xie, et al. Slvit: Scale-wise language-guided vision transformer for referring image segmentation. In *IJCAI*, 2023. 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 2021. 5

[36] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 2021. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention*. Springer, 2015. 1, 2, 5

[38] Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 5

[39] Jaeyoon Sim, Sooyeon Jeon, InJun Choi, et al. Learning to approximate adaptive kernel convolution on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 4

[40] Jaeyoon Sim, Minjae Lee, Guorong Wu, and Won Hwa Kim. Multi-modal graph neural network with transformer-guided adaptive diffusion for preclinical alzheimer classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024. 4

[41] Carole H Sudre, Wenqi Li, Tom Vercauteren, et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *International Workshop on Deep Learning in Medical Image Analysis*. Springer, 2017. 5

[42] Ian Tenney, Patrick Xia, Berlin Chen, et al. What do you learn from context? probing for sentence structure in contextualized word representations. *7th International Conference on Learning Representations, ICLR 2019*, 2019. 2

[43] Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, et al. Tganet: Text-guided attention for improved polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022. 3

[44] Simon Tongbram, Nameirakpam Dhanachandra, and Thelma Ngangom. Hybrid deep learning framework for pancreatic cancer detection and segmentation using graph attention cnn and swin transformer. *ICT Express*, 2025. 1

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. *Advances in neural information processing systems*, 2017. 4

[46] Haonan Wang, Peng Cao, Jiaqi Wang, et al. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *AAAI conference on artificial intelligence*, 2022. 1, 2, 5

[47] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*, 2022. 5

[48] Jianzong Wu, Xiangtai Li, Xia Li, et al. Towards robust referring image segmentation. *IEEE Transactions on Image Processing*, 2024. 5

[49] Renkai Wu, Yinghao Liu, Pengchen Liang, and Qing Chang. H-vmunet: High-order vision mamba unet for medical image segmentation. *Neurocomputing*, 2025. 5

[50] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *Neural Information Processing: 26th International Conference, ICONIP 2019*. Springer, 2019. 4

[51] Zhao Yang, Jiaqi Wang, Yansong Tang, et al. Lavt: Language-aware vision transformer for referring image segmentation. In *IEEE/CVF conference on computer vision and pattern recognition*, 2022. 5

[52] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, et al. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 2019. 4

[53] Yuhao Zhang, Hang Jiang, Yasuhide Miura, et al. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*. PMLR, 2022. 5

[54] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 2018. 5

[55] Yi Zhong, Mengqiu Xu, Kongming Liang, et al. 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. 1, 2, 3, 5, 8

[56] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, et al. 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*. Springer, 2018. 1, 2, 5