HiMix : Hierarchical Visual-Textual Mixing Network for Lesion Segmentation

Soojin Hwang* Jaeyoon Sim* Won Hwa Kim

Medical Information Processing (MIP) Lab, POSTECH

WACV 2026 TUCSON, AZ 3/6 – 3/10

Introduction

- **Objective:** Generating a pixel-wise mask of infected lung regions based on a paired chest image data and its corresponding medical report.
- **Problem:** Existing methods often rely on a single embedding from a text encoder, failing to fully capture the hierarchical structure of linguistic information.
- **Contribution:**
  1. **Hierarchical-Enhanced Visual-Textual Mixing (HiMix)** is a novel multi-modal framework, which refines image and text inputs hierarchically from both modalities.
  2. **Superior performance**: HiMix outperforms SOTA models, proving improvements in both DSC and IoU.
  3. **Adaptability**: Experiments on diverse text formats confirm HiMix's robustness and real-world applicability.

Method - Overview

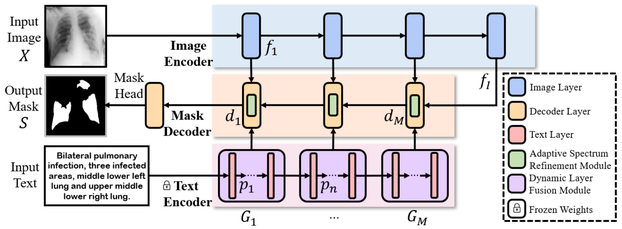

Figure: Architecture of HiMix. An input image X and text prompt are fed into the image and text encoder, respectively. The text feature p is refined through DLFM, while the image feature f is directly fed to the decoder, allowing visual features to be sequentially refined through ASRM to predict the output mask S.

Overview of HiMix

- HiMix consists of three core components including an image encoder, a text encoder, and a mask decoder.
- The model integrates image X and text features across the decoding process to predict the output mask S.
- $\{f_i\}_{i=1}^{I}$: A sequence of features from the input image.
- $\{p_n\}_{n=1}^{N}$: A sequence of features from the input text.
- d_m : A feature of m-th decoder layer.

Method - Key Modules

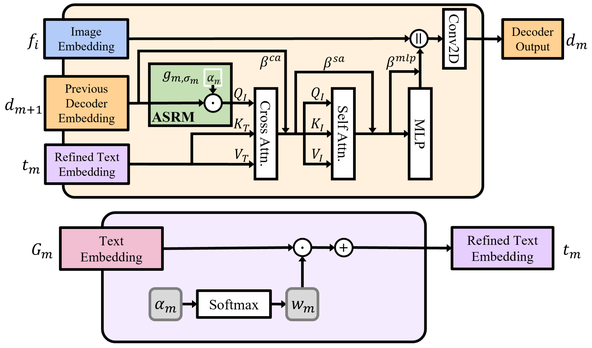

Figure: key modules in HiMix. Up: Decoder with adaptive spectrum refinement module (ASRM). Down: Dynamic layer fusion module (DLFM).

(1) Image Encoder & Text Encoder

- **Dynamic Layer Fusion Module (DLFM)**
  - Text embeddings are split into M groups to match the decoder layers.
  - A fused text feature t_m is constructed from group G_m as

$$t_m = \sum_{p_n \in G_m} w_n p_n.$$

G_m: text group for decoder layer m

w_n : softmax weight within group

(2) Mask Decoder

- **Adaptive Spectrum Refinement Module (ASRM)**
  - Frequency-domain Gaussian filtering to enhance global-local visual context
  - A refined visual feature $\hat{d}_m$ is computed as

$$\hat{d}_m = \mathrm{IFFT}(\mathrm{FFT}(d_m) \odot g_m) + d_m.$$

$\odot$: An element-wise multiplication

g_m : A Gaussian filter with learnable bandwidth

- Final segmentation mask S is produced by a segmentation head.

Experiments Result

(1) Quantitative Comparisons

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

(2) Qualitative Comparisons

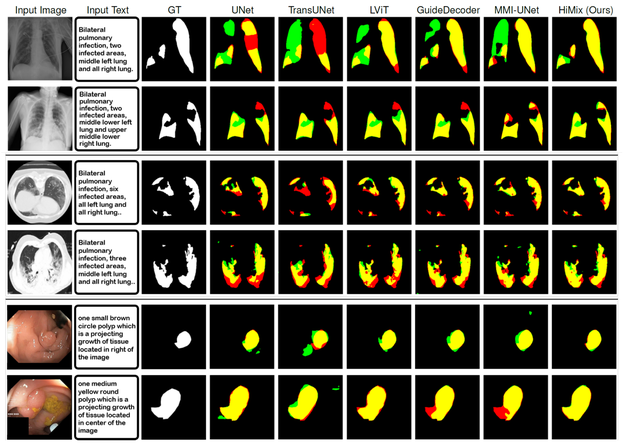

Figure: Visualization of segmentation results. Results on the QaTa-COV19 (top), MosMedData+ (middle), and Kvasir-SEG (bottom) datasets. Yellow, red, and green represent true positive, false negative, and false positive, respectively.

Additional Analysis

(1) Ablation Studies on Key Modules/Text Encoders

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

Module ASRM	Module DLFM	QATA DSC ↑	QATA IoU ↑	MOSMED DSC ↑	MOSMED IoU ↑	KVASIR DSC ↑	KVASIR 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**

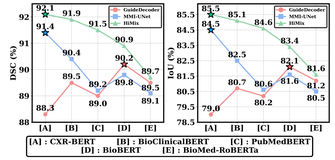

Figure: Ablation study of text encoders on the KVASIR dataset. The best result for each method is marked with a star, and across all text encoders, HiMix consistently achieves the best performance.

(2) Analysis of Spectrum Refinement/Anatomical Relationships

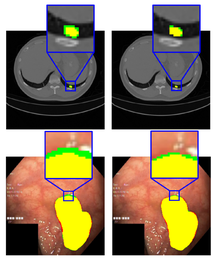

Figure: Qualitative comparison on the MOSMED (top) and KVASIR (Bottom), showing results without ASRM (left) and with ASRM (right) for each dataset.

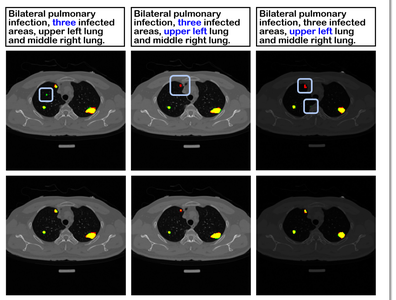

Figure: Ablation study on the importance of anatomical relationships in text on the MOSMED. A comparison of segmentation predictions from a model that uses only the final layer of the text encoder (top) and HiMix (bottom).

(3) Generalization of HiMix

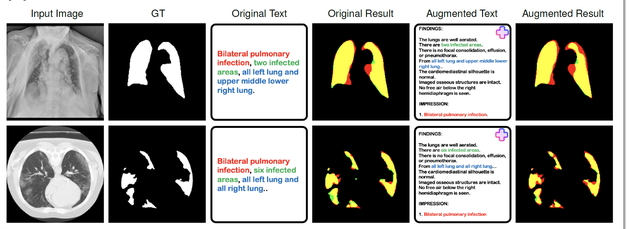

Figure: Effect of text augmentation for segmentation. The original text is structured, while the augmented text mimics medical reports from MIMIC-CXR.