HiMix : Hierarchical Visual-Textual Mixing Network for Lesion Segmentation (Supplementary Material)

This material presents the supplementary paper from the main paper due to space limitations, providing additional analyses and visualizations that complement the main manuscript.

1. Comparison to Recent Language-Guided Segmentation Methods

To better highlight the architectural novelty of our model, Fig. 1 compares representative designs of prior multi-modal segmentation methods.

- In (a), early fusion approaches such as LViT [6] combine visual and textual features at the encoder level. However, such design limits the influence of language to the inital stages of feature extraction, failing to propagate semantic cues throughout the entire network.
- (b) shows methods such as GuidedDecoder [10], which introduce text guidance in the decoding stage using a fixed embedding. These approach typically use a single fixed embedding to condition all decoder stages uniformly, thereby underutilizing the granularity and semantic detail of text information.
- (c) depicts MMI-UNet [1], which jointly fuses image and text features in the decoder. While this approach introduces a deeper interaction than the above methods, it still lack explicit multi-level alignment and progressive refinement across hierarchical layers.
- In contrast, (d) illustrates our approach, which explicitly refines and aligns multi-level visual and textual features at each decoding stage through our proposed modules. This stage-wise guidance mechanism allows our model to retain both the global semantics and the fine-grained spatial alignment encoded in the input text. Moreover, unlike prior works that statically fuse language once, our architecture hierarchically propagates and adapts textual semantics throughout the entire decoding process, ensuring a richer and more contextually aligned segmentation.

2. Dataset Description

Three representative benchmarks, i.e., Qata-COV19, MosMedData+ and Kvasir-SEG, are used in the experiments for lesion segmentation to evaluate the performance of our method.

Qata-COV19 dataset. The Qata-COV19 dataset [3] is compiled by researchers from Qatar University and Tampere University. This dataset consists of 9258 COVID-19 chest X-ray radiographs with manual annotations of COVID-19 lesions for the first time. [6] extends text annotations on the QaTa-COV19 dataset with the help of medical experts. The text annotations focus on whether both lungs are infected, the number of lesion regions, and the approximate location of the infected areas. For example, "Bilateral pulmonary infection, two infected areas, upper left lung and upper right lung." refers to bilateral lung infection, and there are two infection areas located in the upper left lung and the upper right lung respectively.

MosMedData+ dataset. The MosMedData+ dataset [8] contains anonymised human lung Computed Tomography (CT) scans with COVID-19 related findings, as well as without such finding. CT scans were obtained between 1st of March, 2020 and 25th of April, 2020, and provided by municipal hospitals in Moscow, Russia. This dataset includes 2729 CT scan slices of lung infections. The text annotations on MosMedData+ dataset mainly contain the same information as QaTa-COV19 dataset.

Kvasir-SEG dataset. The Kvasir-SEG dataset [4] is an open-access dataset of gastrointestinal polyp images and corresponding segmentation masks, manually annotated by a medical doctor. Specifically, these annotations have been meticulously crafted and validated by an experienced gastroenterologist. This dataset consists of 1000 high-quality gastrointestinal polyp images. For example, "One small bron circle polyp which is a projecting growth of tissue located in right of the image." refers to a small round polyp, which is a projecting growth of tissue, located on the right side of the image.

3. Implementation Details

Our proposed method is implemented using Pytorch. Training and testing were performed on a single NVIDIA RTX A6000 GPU with 48GB of memory. For optimization, we employed a combination of Dice loss and Cross-Entropy loss using the AdamW [7] optimizer with a batch size of 32. The learning rate follows a cosine annealing schedule,

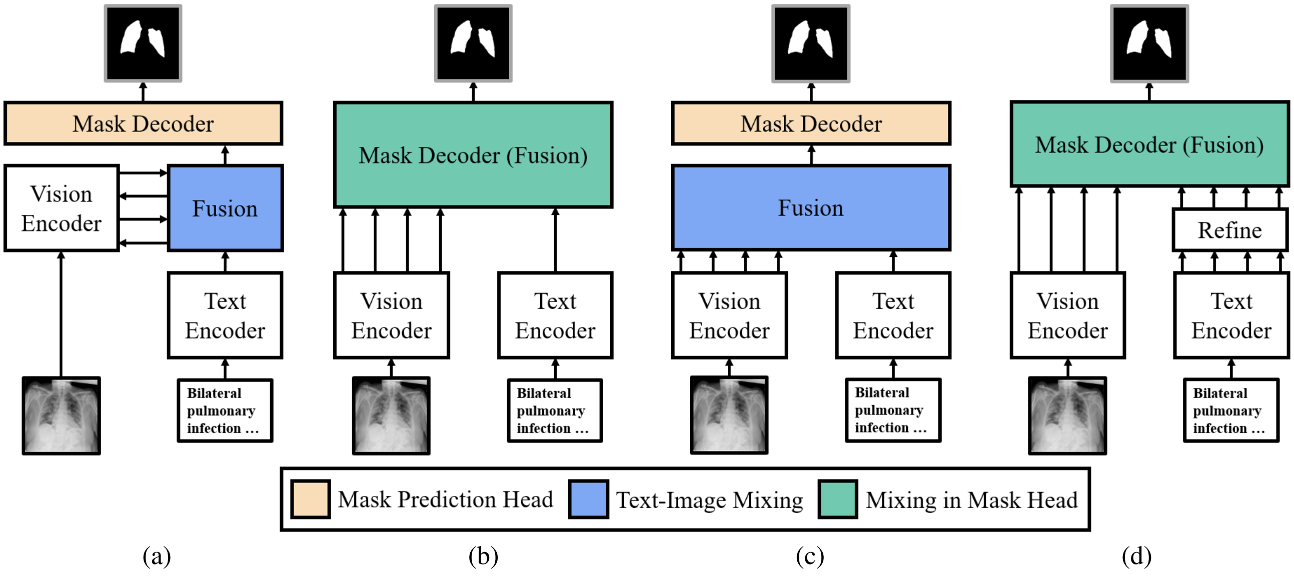


Figure 1. A comparative analysis of different model architectures for mask modeling and language-guided medical image segmentation.

starting at 3e-4 and decreasing to 1e-6. We also use an early stop mechanism until the performance of model does not improve for 20 epochs. In DLFM, the weights for text embeddings are initialized to 1, allowing the model to learn optimal layer-wise importance during training. For ASRM, the initial kernel bandwidth of the Gaussian filtering coefficient map at each decoder layer is set to 5 and jointly optimized in an end-to-end manner.

For training, we used composite loss function L_{total} that combines Dice loss L_{Dice} and Cross-Entropy loss L_{CE} as

$$L_{\text{Dice}} = 1 - \frac{1}{NC}\sum_{i=1}^{N}\sum_{j=1}^{C}\frac{2|p_{ij} \cap y_{ij}|}{|p_{ij}| + |y_{ij}|}$$

$$L_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(p_i j)$$

$$L_{\text{total}} = \frac{1}{2}(L_{\text{Dice}} + L_{\text{CE}})$$

where N represents the number of pixels, C denotes the number of categories, which is set to 1 in our experiments. Here, p_{ij} is the prediction probability that pixel i belongs to category j, y_{ij} indicates whether pixel i belongs to category j. If pixel i belongs to category j, then y_{ij} is 1, otherwise 0. This formulation balances region-level overlap and pixel-wise classification accuracy, providing stable optimization during training.

The text annotations on the medical segmentation benchmarks, i.e., QaTa-COV19, MosMedData+, and Kvasir-SEG datasets, are extended with the help of professionals. For example, the text annotations for pulmonary lesions focus on whether both lungs are infected, the number of lesion regions, and the approximate location of the infected areas. The radiologists independently annotated the same image, and then the authors from [6] compared their annotations to ensure consistency. Additionally, they conducted a quality check based on the provided mask to ensure that there was no excessive deviation in the text annotations. Before training, using the tokenizer from the CXR-BERT, we preprocess text descriptions for the text encoder. Each text description is tokenized and then padded to a maximum length of 24. This ensures that all text sequences have a consistent length, allowing them to be correctly processed by the text encoder.

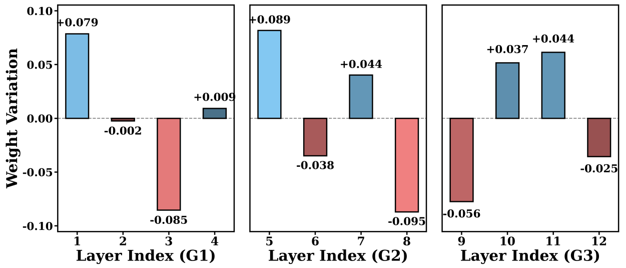


Figure 2. Variation of w values across text embedding layers for (G_3, G_2, G_1) on the MOSMED dataset. Final embedding shows reduced impact despite varied layer-wise contributions.

4. Analysis of Layer-wise Contribution

As shown in Fig. 2, we explored the changes in text embedding contributions after training by uniformly initializing before optimization. Contrary to prior assumptions [1, 10], the embedding weight of the final layer decreases by 0.025,

Table 1. Performance comparison with different DLFM grouping orders on both QATA and MOSMED datasets.

DLFM Order	QaTa-COV19		MosMedData+	
	DSC ↑	**IoU** ↑	**DSC** ↑	**IoU** ↑
no grouping	90.90	83.31	78.36	64.42
(G_1,G_2,G_3)	91.04	83.55	78.46	64.55
(G_1,G_3,G_2)	90.99	83.47	78.18	64.18
(G_2,G_1,G_3)	90.92	83.36	77.80	63.66
(G_2,G_3,G_1)	90.96	83.42	78.82	65.04
(G_3,G_1,G_2)	90.81	83.17	78.09	64.05
(G_3,G_2,G_1)	**91.17**	**83.78**	**79.44**	**65.90**

Table 2. Top: Comparison of grouping strategies in DLFM. Performance is reported on the MOSMED dataset. All variants use the same DLFM module; only grouping differs. Bottom: Ablation study on the Text Encoder groups. Performance of HiMix when individually removing one of the three Text Encoder groups (G1, G2, G3) are compared.

Grouping Method	Mean Pooling	Interleaved	Clustering	Ours
DSC ↑	78.15	78.55	78.09	79.44
IOU ↑	64.14	64.68	64.06	65.90
Text Encoder Remove	w/o G1	w/o G2	w/o G3	Ours
DSC ↑	78.41	77.50	78.07	79.44
IOU ↑	64.49	63.26	64.03	65.90

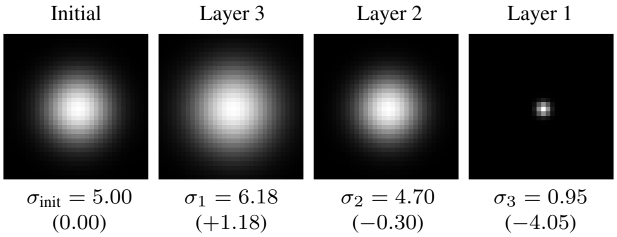


Figure 3. Variation of filtering coefficient map (top) and trained σ values (bottom) across decoder layers on the MOSMED dataset.

showing that relying solely on it does not improve segmentation. Instead, adaptively distributing weight across all embeddings enhances performance, emphasizing the role of ordering in medical text-driven segmentation.

5. Effect of Text Embedding Sequence

To validate DLFM, we examined how the order of fusing text embeddings affects segmentation. In Table 1, we split 12 text embeddings into three consecutive sets, each consisting of four embeddings aligned with one decoder layer. We then tested all possible group sequences by permuting the order of them. Without grouping, all 12 embeddings were equally provided to all decoder layers, rather than being divided and sequentially fused. Although the naive strategy achieves reasonable performance, our hierarchical grouping approach, particularly the (G_3, G_2, G_1) sequence, yields the best DSC and IoU scores, with values of 91.17 and 83.78 on QATA, and 79.44 and 65.90 on QATA, respectively. This alignment of embedding order with the image encoder hierarchy suggests that incorporating high-level semantics early and fine-grained details later is more effective.

6. Additional Ablations on DLFM Grouping

Ablation Study on Grouping Method. To further validate the robustness of DLFM, we investigated several alternative strategies for grouping the embeddings from the text encoder layer, which are reported in the top of Table. 2. Let the encoder provide N hidden states (in our case, CXR-BERT with $N = 12$). These were partitioned into K groups ($K = 3$ in our experiments, corresponding to the three decoder stages). We compared the following strategies:

- **Original.** The N layers are divided into K sequential blocks, e.g., $\{1..N/K\}, \{N/K+1..2N/K\}, ...$, and each group is aggregated with learnable softmax weights.
- **Mean pooling.** Same sequential partition, but each group embedding is the mean of its member layers.
- **Interleaved.** Layers are distributed to groups in a round-robin fashion with stride K, e.g., $\{1, 1 + K, 1 + 2K, ...\}, \{2, 2 + K, ...\}, ...$. Learnable softmax weighting is applied within each group.
- **Clustering.** Groups are obtained by applying k-means clustering to layer-wise proxy statistics (mean and variance of token embeddings, reduced with PCA). Groups are then ordered by their average layer index (high→coarse, mid→middle, low→fine) to align with decoder stages.

Ablation Study on Text Encoder Groups. In the bottom of Table 2, it shows the results of an ablation study on the Text Encoder groups, which highlights the importance of its intermediate layers for segmentation performance. The study examined the impact of removing each of the three groups (i.e., G1, G2, and G3) from the Text Encoder on the segmentation performance. The results indicate that the absence of G2 leads to the most significant performance degradation, with DSC dropping to 77.50% and IoU to 63.26%. This finding supports our hypothesis that a hierarchical structure requires information from intermediate layers to function optimally. By capturing and utilizing the multi-level information from the intermediate layers, the model is able to more effectively guide the segmentation process, leading to an improvement in overall performance.

7. Analysis of Spectrum Refinement

Fig. 3 illustrates how the Gaussian filtering coefficient map evolves across decoder layers, with learnable σ values representing the kernel bandwidth at each stage. All decoder stages are set with a kernel bandwidth of σ_{init}=5.00. Dur-

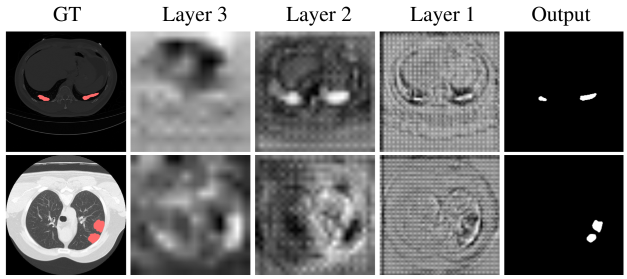


Figure 4. Visualization of segmentation outputs from decoder layers on the MOSMED dataset. Each column shows the input image with lesions, outputs at three stages, and the final prediction mask. HiMix effectively captures global-to-local lesion features. Trained σ values across decoder layers are on the bottom.

ing training, the first decoder layer (i.e., Layer 3) adaptively increases σ to 6.18, promoting the capture of global context. In contrast, the later decoder layers (i.e., Layer 2 and Layer 1) reduce their σ to 4.70 and 0.95, respectively, shifting focus toward localized details. This stage-specific bandwidth pattern shows that the learned σ values guide the decoder from global to local representation, with larger σ in early layers capturing global context (e.g., disease distribution) and smaller σ in later layers emphasizing local details (e.g., lesion margins).

Building on the optimized kernel band-widths in Fig. 3, Fig. 4 illustrates decoder outputs across stages of mask prediction. As decoding progresses, the receptive fields, modulated by optimized kernel scales, shift from capturing coarse structures (i.e., low-frequency components with larger σ) to fine-grained details (i.e., high-frequency components with smaller σ). This transition yields increasingly precise delineation of lesion boundaries at deeper stages. Notably, early decoder outputs capture the global lesion shape by incorporating broad contextual cues, whereas later stages enhance boundary sharpness by focusing on localized anatomical features. This hierarchical refinement reveals that HiMix effectively leverages frequency-specific information at each stage, gradually improving spatial precision without compromising semantic coherence. Such spectrum-aware decoding contributes to both robust and interpretable segmentation performance.

8. Statistical validation of improvements

Since all prior baselines use the fixed train, validation, and test split, Table 1 in the main paper reports mean performance to ensure reproducible and fair comparisons. To capture variability, Table 3 reports DSC and IoU as mean and standard deviation. Compared with MedSAM, HiMix shows larger standard deviations, yet it delivers substantially higher mean DSC and IoU. In contrast, against the prior SOTA method, i.e., MMI-UNet, HiMix exhibits consistently smaller standard deviations across datasets and metrics, indicating robustness and superiority of HiMix.

Table 3. Statistical validation of improvements on the various benchmarks.

Method	QATA		MOSMED	
	DSC ↑	IoU ↑	DSC ↑	IoU ↑
MedSAM	$78.49_{\pm 2.24}$	$69.11_{\pm 2.50}$	$54.22_{\pm 2.89}$	$42.22_{\pm 2.57}$
MMI-UNet	$90.88_{\pm 3.48}$	$83.28_{\pm 5.64}$	$78.42_{\pm 6.58}$	$64.50_{\pm 7.98}$
HiMix(Ours)	$91.17_{\pm 3.24}$	$83.78_{\pm 4.74}$	$79.44_{\pm 5.92}$	$65.90_{\pm 7.23}$

9. Additional Qualitative Results

To further evaluate the qualitative performance of our model, we present extended visual comparisons in Fig. 5 and 6, covering both the QaTa-COV19 and MosMedData+ datasets. Each row shows the segmentation results from various baseline models, including UNet [9], TransUNet [2], LViT [6], GuideDecoder [10], and MMI-UNet [1], alongside our proposed method. Ground truth masks and corresponding input image-text pairs are also provided for reference. Consistent with the results discussed in the main paper, HiMix demonstrates robust and accurate lesion segmentation across diverse cases. In particular, it yields fewer false positives and false negatives (represented by green and red, respectively) while preserving high true positive coverage (yellow). This improvement is especially evident in complex or subtle infection regions, highlighting the ability of HiMix to effectively align spatial and textual cues. These findings support the advantage of our language-guided segmentation framework in both structured and realistic clinical settings.

10. Hierarchical Segmentation Visualization across Decoder Stages

To provide an intuitive understanding of the progressive refinement of segmentation masks through different decoder layers, we visualize the outputs from intermediate decoder stages across two datasets, Qata-COV19 and MosMed-Data+. As shown in Fig. 7 and Fig. 8, each row represents the segmentation precess for a single input image. From left to right, we display the input image, the predicted masks from decoder layers 1 through 3, and the final segmentation output. We observe that earlier decoder layers (e.g., decoder layer 1) produce coarse and blurry activations, while deeper decoder layers (e.g., decoder layer 3) progressively sharpen the lesion boundaries and suppress irrelevant regions. Notably, decoder layer 3 often captures lesion contours with enhanced clarity, even before the final prediction layer refines the output. These results support our assumption in ASRM that finer decoder stages contribute critical spatial details. The multi-stage refinement design enables the model to leverage both coarse semantic features and high-resolution details, which are essential for accurate lesion localization in challenging medical images.

11. Evaluation of Robustness

To complement the analysis presented in the main paper, we provide additional qualitative results for Qata-COV19 in Fig. 9 and for MosMedData+ in Fig. 10. Due to space limitations, these examples were not included in the main text. They illustrate how our model generalizes to complex and realistic report-style text inputs. While our model is trained solely on structured prompt sentences, it consistently maintains accurate segmentation even when tested with augmented text that mimics free-form clinical reports from MIMIC-CXR [5]. This suggests that HiMix captures core semantic cues and does not overly rely on rigid syntactic structures. The augmented results show comparable performance to structured text inputs in terms of capturing infected regions, highlighting the robustness of our language-guided method in real-world scenarios.

12. Limitations and Future Directions

While our model demonstrates strong performance in both structured and augmented text scenarios, certain limitations remain. First, the framework relies on the presence of descriptive input text to guide segmentation, which may not always be available or may vary in quality across clinical settings. Although our experiments show robustness to textual variation, extreme cases with ambiguous or contradictory language may still affect prediction reliability. Second, our current design processes each image-text pair independently, without incorporating temporal context or inter-slice correlations that could further enhance consistency in volumetric data. Nonetheless, these limitations arise because our model is designed to align language and visual features precisely at each image slice, which in turn helps it generalize well to new and unfamiliar text styles. Future work could explore integrating temporal context or self-refinement mechanisms to further improve stability in longitudinal or sequential analyses.

References

[1] 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*, pages 702–711. Springer, 2024. 1, 2, 4

[2] Jieneng Chen, Yongyi Lu, Qihang Yu, et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 4

[3] 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. 1

[4] Debesh Jha, Pia H Smedsrud, Michael A Riegler, et al. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019. 1

[5] 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*, 6 (1):317, 2019. 5

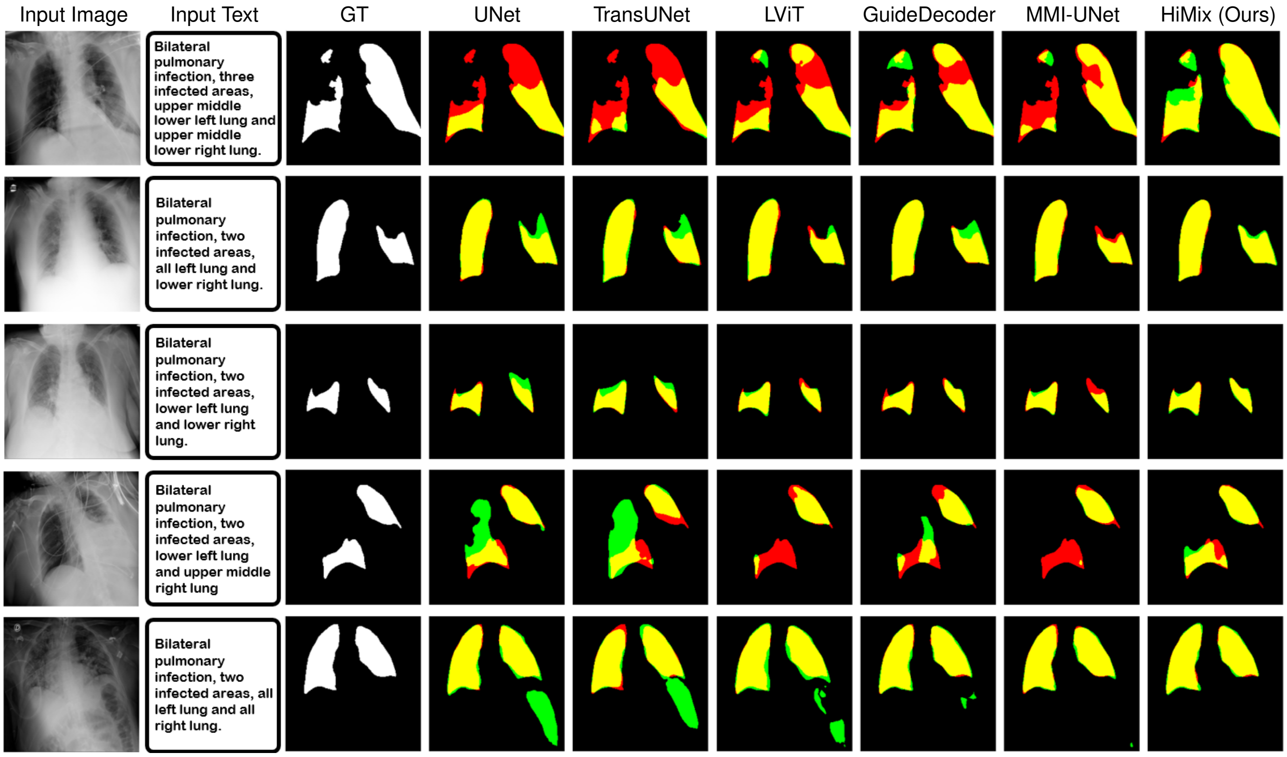
[6] Zihan Li, Yunxiang Li, Qingde Li, et al. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 43(1):96–107, 2023. 1, 2, 4

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2017. 1

[8] Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, et al. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *medRxiv*, 2020. 1

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015*, pages 234–241. Springer, 2015. 4

[10] 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*, pages 724–733. Springer, 2023. 1, 2, 4



Bilateral pulmonary infection, three infected areas, upper middle lower left lung and upper middle lower right lung.

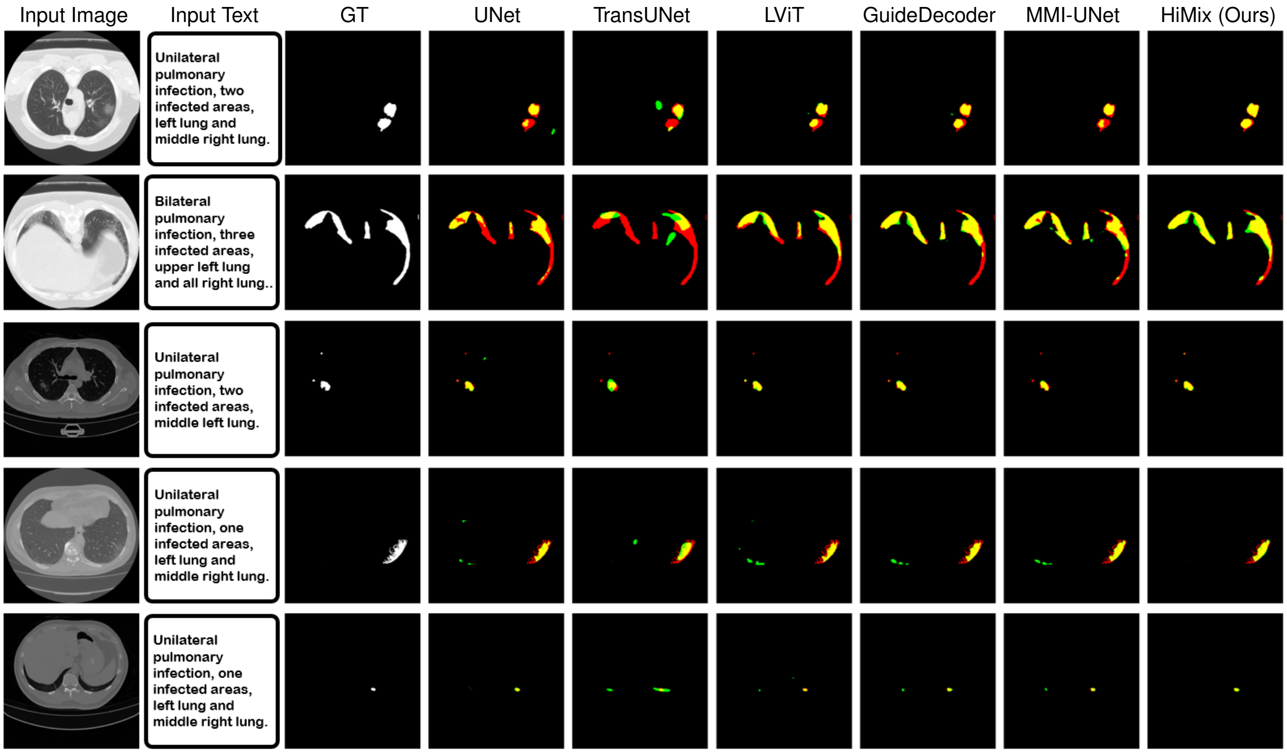
Bilateral pulmonary infection, two infected areas, all left lung and lower right lung.

Bilateral pulmonary infection, two infected areas, lower left lung and lower right lung.

Bilateral pulmonary infection, two infected areas, lower left lung and upper middle right lung

Bilateral pulmonary infection, two infected areas, all left lung and all right lung.

Figure 5. **Visualization of segmentation results.** Results on QaTa-COV19 dataset. Yellow, red, and green represent true positive, false negative, and false positive, respectively.



Unilateral pulmonary infection, two infected areas, left lung and middle right lung.

Bilateral pulmonary infection, three infected areas, upper left lung and all right lung..

Unilateral pulmonary infection, two infected areas, middle left lung.

Unilateral pulmonary infection, one infected areas, left lung and middle right lung.

Unilateral pulmonary infection, one infected areas, left lung and middle right lung.

Figure 6. **Visualization of segmentation results.** Results on MosMedData+ dataset. Yellow, red, and green represent true positive, false negative, and false positive, respectively.

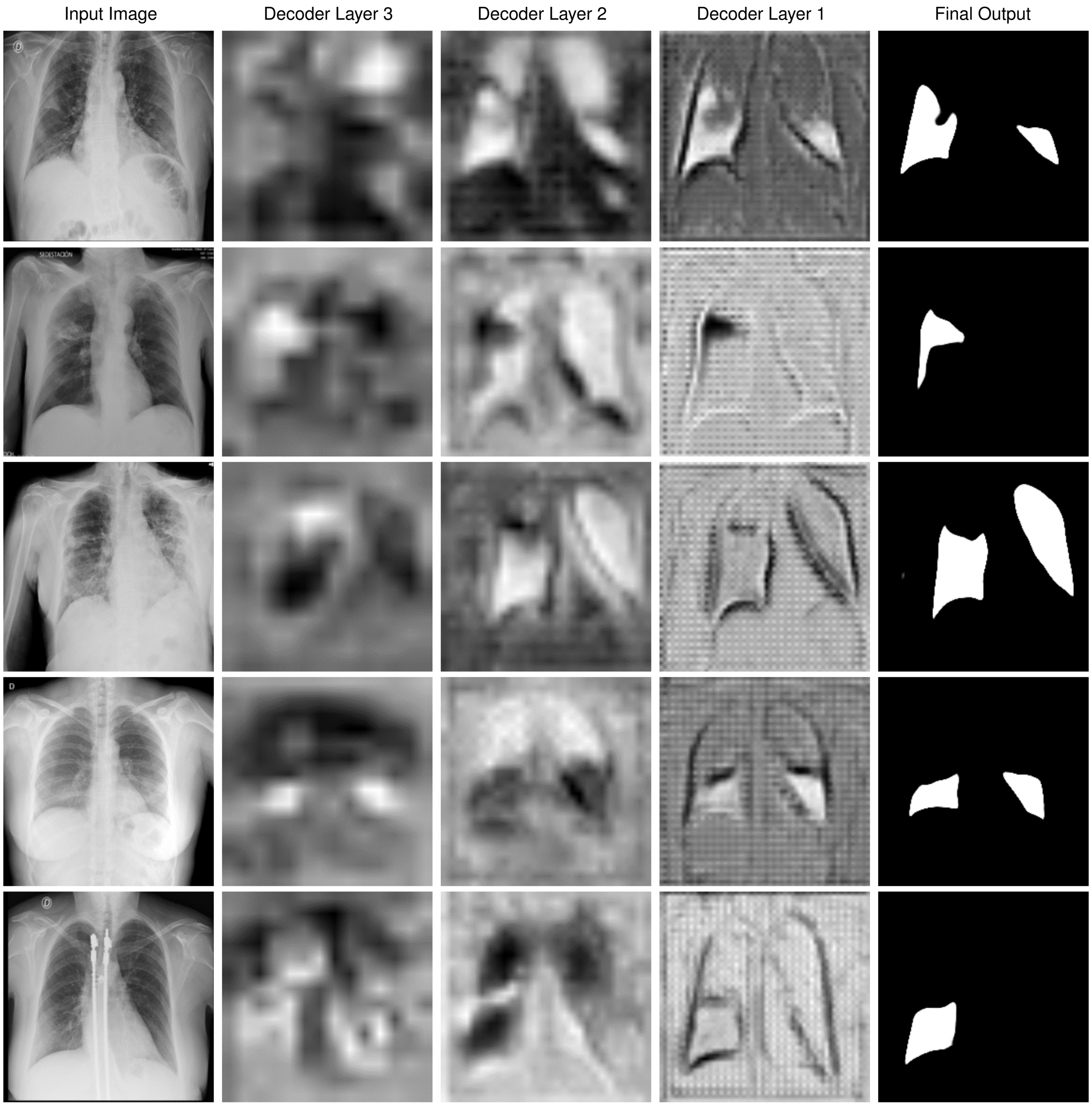


Figure 7. **Visualization of segmentation outputs from intermediate decoder layers on the Qata-COV19 dataset.** Each column shows the input image, decoder outputs at three stages, and the final prediction mask. Segmentation maps become progressively refined and spatially accurate through deeper layers.

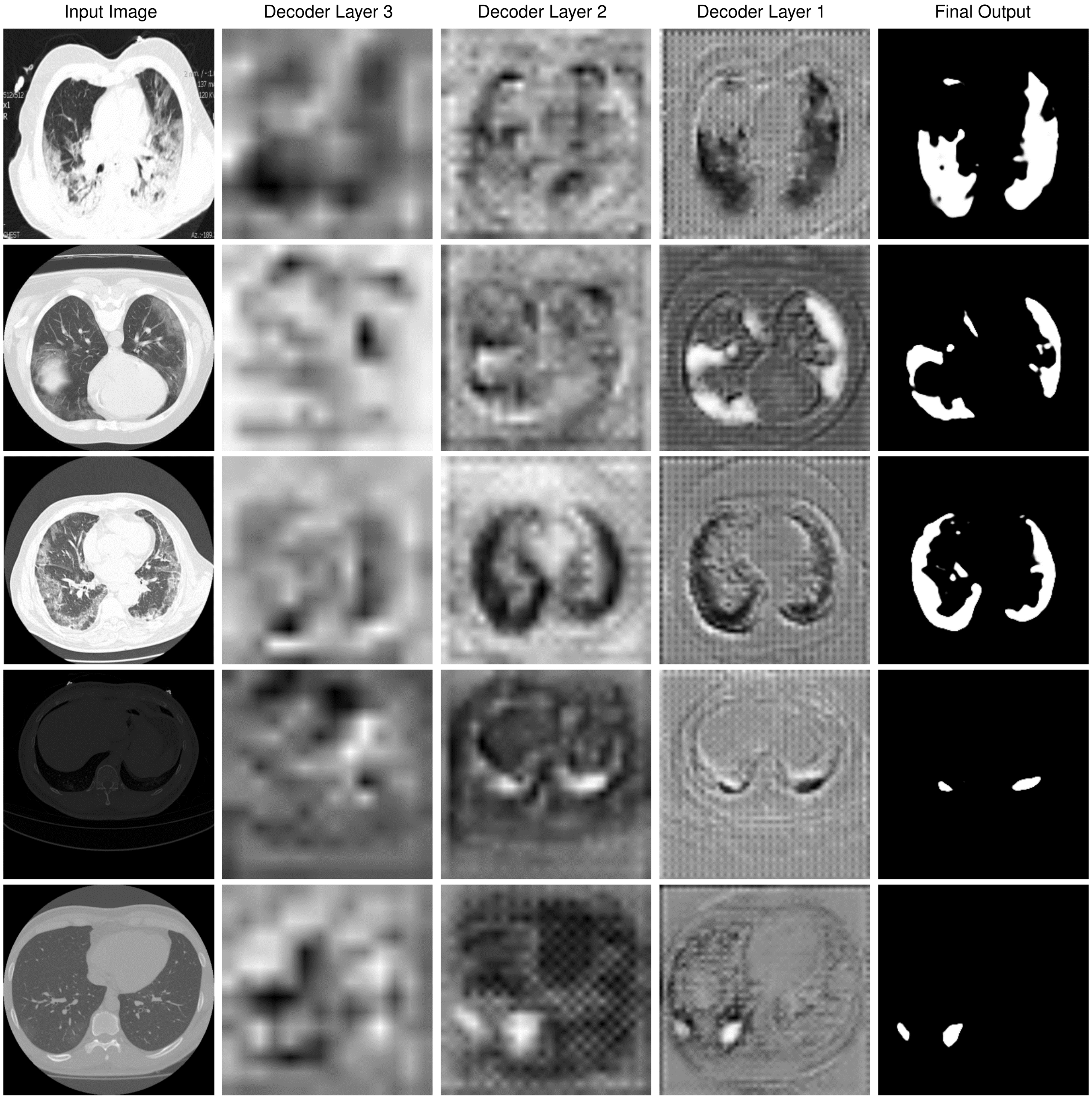
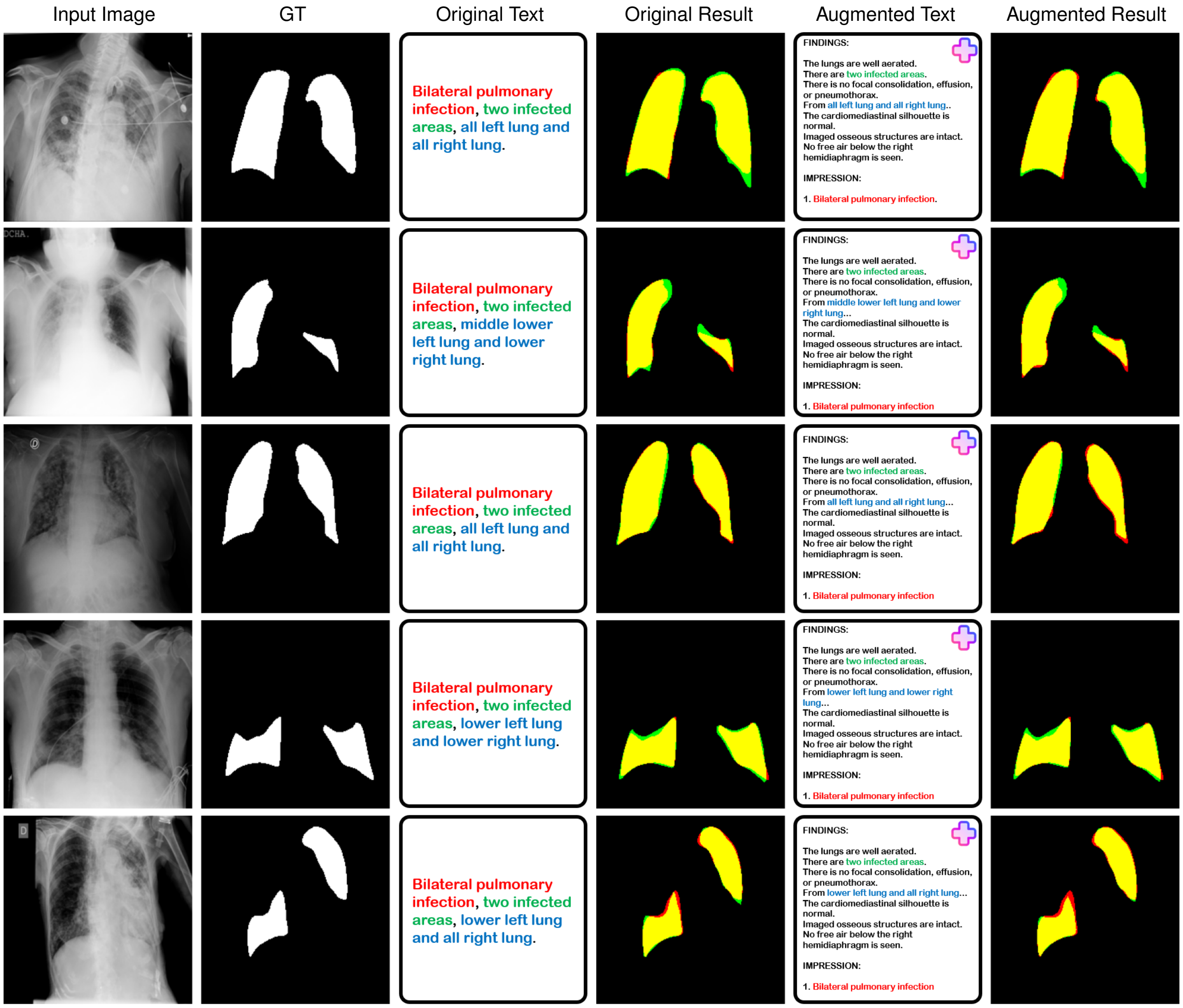


Figure 8. **Visualization of segmentation outputs from intermediate decoder layers on the MosMedData+ dataset.** Each column shows the input image, decoder outputs at three stages, and the final prediction mask. Segmentation maps become progressively refined and spatially accurate through deeper layers.



Bilateral pulmonary infection, two infected areas, all left lung and all right lung.

FINDINGS:

The lungs are well aerated.
There are two infected areas.
There is no focal consolidation, effusion, or pneumothorax.
From all left lung and all right lung..
The cardiomediastinal silhouette is normal.
Imaged osseous structures are intact.
No free air below the right hemidiaphragm is seen.

IMPRESSION:

1. Bilateral pulmonary infection.

Bilateral pulmonary infection, two infected areas, middle lower left lung and lower right lung.

FINDINGS:

The lungs are well aerated.
There are two infected areas.
There is no focal consolidation, effusion, or pneumothorax.
From middle lower left lung and lower right lung...
The cardiomediastinal silhouette is normal.
Imaged osseous structures are intact.
No free air below the right hemidiaphragm is seen.

IMPRESSION:

1. Bilateral pulmonary infection

Bilateral pulmonary infection, two infected areas, all left lung and all right lung.

FINDINGS:

The lungs are well aerated.
There are two infected areas.
There is no focal consolidation, effusion, or pneumothorax.
From all left lung and all right lung...
The cardiomediastinal silhouette is normal.
Imaged osseous structures are intact.
No free air below the right hemidiaphragm is seen.

IMPRESSION:

1. Bilateral pulmonary infection

Bilateral pulmonary infection, two infected areas, lower left lung and lower right lung.

FINDINGS:

The lungs are well aerated.
There are two infected areas.
There is no focal consolidation, effusion, or pneumothorax.
From lower left lung and lower right lung...
The cardiomediastinal silhouette is normal.
Imaged osseous structures are intact.
No free air below the right hemidiaphragm is seen.

IMPRESSION:

1. Bilateral pulmonary infection

Bilateral pulmonary infection, two infected areas, lower left lung and all right lung.

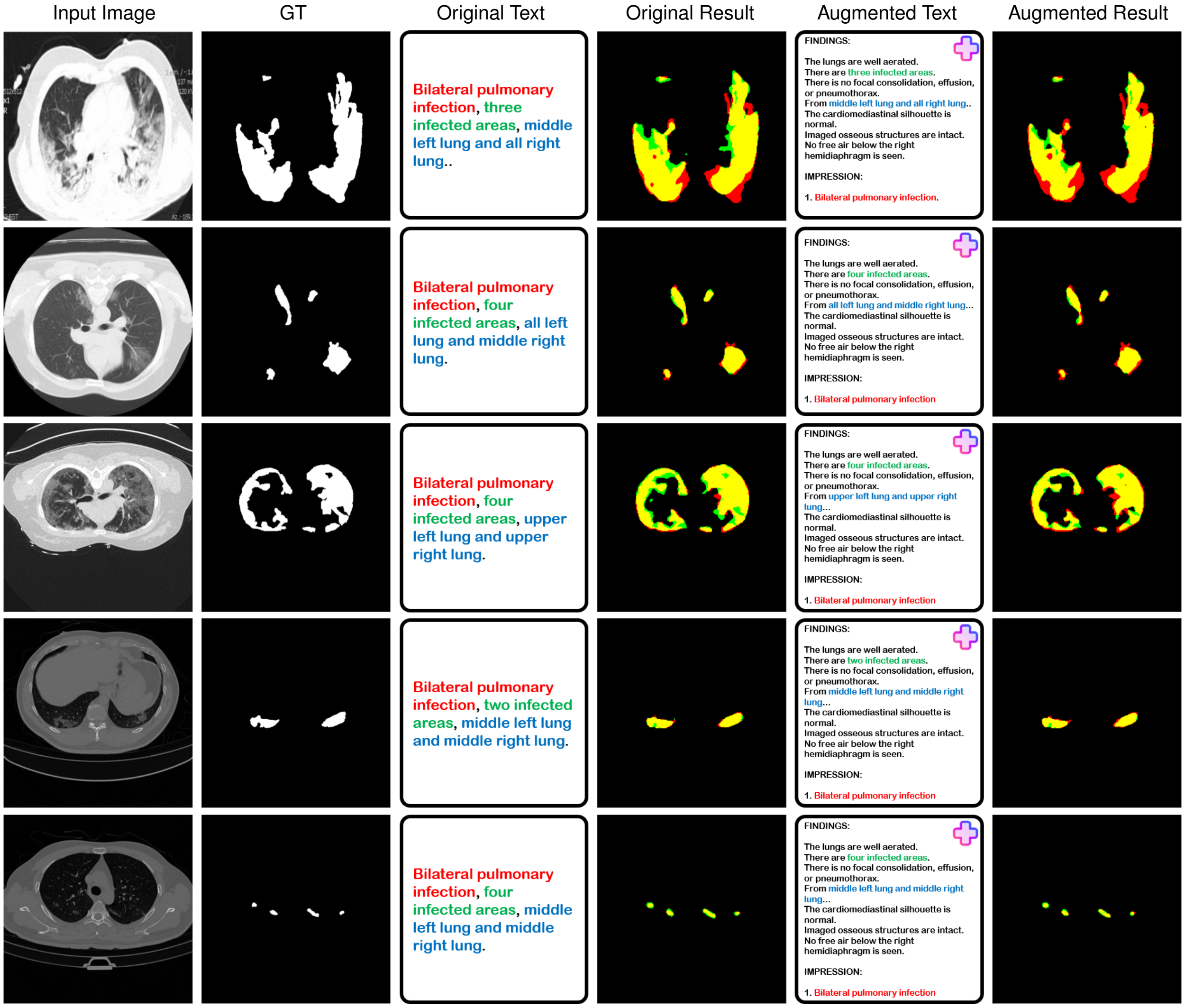
FINDINGS:

The lungs are well aerated.
There are two infected areas.
There is no focal consolidation, effusion, or pneumothorax.
From lower left lung and all right lung...
The cardiomediastinal silhouette is normal.
Imaged osseous structures are intact.
No free air below the right hemidiaphragm is seen.

IMPRESSION:

1. Bilateral pulmonary infection

Figure 9. **Effect of text augmentation for segmentation on QaTA-COV19 dataset.** The original text is structured, while the augmented text mimics medical reports from MIMIC-CXR.



Bilateral pulmonary infection, three infected areas, middle left lung and all right lung..

FINDINGS:

The lungs are well aerated.
There are three infected areas.
There is no focal consolidation, effusion, or pneumothorax.
From middle left lung and all right lung..
The cardiomediastinal silhouette is normal.
Imaged osseous structures are intact.
No free air below the right hemidiaphragm is seen.

IMPRESSION:

1. Bilateral pulmonary infection.

Bilateral pulmonary infection, four infected areas, all left lung and middle right lung.

FINDINGS:

The lungs are well aerated.
There are four infected areas.
There is no focal consolidation, effusion, or pneumothorax.
From all left lung and middle right lung...
The cardiomediastinal silhouette is normal.
Imaged osseous structures are intact.
No free air below the right hemidiaphragm is seen.

IMPRESSION:

1. Bilateral pulmonary infection

Bilateral pulmonary infection, four infected areas, upper left lung and upper right lung.

FINDINGS:

The lungs are well aerated.
There are four infected areas.
There is no focal consolidation, effusion, or pneumothorax.
From upper left lung and upper right lung...
The cardiomediastinal silhouette is normal.
Imaged osseous structures are intact.
No free air below the right hemidiaphragm is seen.

IMPRESSION:

1. Bilateral pulmonary infection

Bilateral pulmonary infection, two infected areas, middle left lung and middle right lung.

FINDINGS:

The lungs are well aerated.
There are two infected areas.
There is no focal consolidation, effusion, or pneumothorax.
From middle left lung and middle right lung...
The cardiomediastinal silhouette is normal.
Imaged osseous structures are intact.
No free air below the right hemidiaphragm is seen.

IMPRESSION:

1. Bilateral pulmonary infection

Bilateral pulmonary infection, four infected areas, middle left lung and middle right lung.

FINDINGS:

The lungs are well aerated.
There are four infected areas.
There is no focal consolidation, effusion, or pneumothorax.
From middle left lung and middle right lung...
The cardiomediastinal silhouette is normal.
Imaged osseous structures are intact.
No free air below the right hemidiaphragm is seen.

IMPRESSION:

1. Bilateral pulmonary infection

Figure 10. **Effect of text augmentation for segmentation on MosMedData+.** The original text is structured, while the augmented text mimics medical reports from MIMIC-CXR.