# Learning Role-Conditioned Alignment for Medical Image Referring Segmentation

Kun Li[1], Fu Wang[1], Feixiang Zhou[1], He Zhao[1], Xiaowei Xu[2], Yanda Meng[3], Yitian Zhao[4], and Yalin Zheng[1,⋆]

[1] University of Liverpool, Liverpool, UK
yalin.zheng@liverpool.ac.uk
[2] Leibniz Institute for Analytical Sciences, Dortmund, Germany
[3] King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
[4] Chinese Academy of Sciences, Ningbo, China

**Abstract.** Accurate segmentation of infected regions in radiological imagery is fundamental for the diagnosis and treatment of pulmonary infectious diseases. While recent advancements in multimodal learning have introduced clinical reports as textual guidance for medical image segmentation, existing early- and late-fusion frameworks predominantly treat the textual input as a monolithic global feature. Such approaches neglect the distinctive semantic roles of individual text tokens–specifically infection type, quantity, and location–resulting in diffuse, non-specific guidance. This lack of granularity leads to feature redundancy and suboptimal performance in complex clinical scenarios where diverse lesion morphologies require precise grounding. To address these limitations, we propose a novel **R**ole-**C**onditioned **A**lignment model (**RCA**) for medical image referring segmentation. RCA effectively integrates text-guided cues into visual features by explicitly exploiting the unique functional characteristics of different text tokens to enhance cross-modal feature alignment. Furthermore, we introduce a Boundary Refinement Module (BRM) that leverages edge-aware information to precisely delineate fuzzy boundaries where lesions overlap with healthy tissue. Extensive experiments conducted on the QaTa-COV19 and MosMedData+ benchmarks demonstrate the effectiveness of our proposed method, which achieves state-of-the-art results and significantly outperforms current unimodal and multimodal approaches. The code is publicly available at https://github.com/lik1996/RCA.

**Keywords:** Medical Image Segmentation · Multimodal Learning.

## 1 Introduction

Medical image segmentation (MIS) serves as a critical prerequisite for clinical decision-making, providing the quantitative foundation for computer-aided diagnosis (CAD), treatment planning, and longitudinal monitoring [1]. Recent

⋆ Corresponding author

advances in deep learning, particularly Convolutional Neural Networks (CNNs [15,23]) and Transformers [4,24], have markedly improved automated delineation in medical imaging. Despite this progress, pathological lesions often induce severe structural distortions and intensity fluctuations that can degrade the performance. Furthermore, the intricate morphological characteristics and fuzzy boundaries of diseased regions frequently overlap with healthy tissue, leading to semantic ambiguity that visual features alone struggle to resolve.

To overcome these vision-centric limitations, **Medical Image Referring Segmentation (MIRS)** has emerged as a promising paradigm. Drawing inspiration from referring segmentation in natural images [18,26], MIRS leverages natural language as a guide to provide semantic context, facilitating the distinction of targets from complex background. In clinical practice, pulmonary imaging is typically accompanied by expert reports that describe lesion morphology, spatial distribution, and quantity. By integrating these high-level semantic priors–such as infection type and location–models can effectively disambiguate complex clinical scenes. To facilitate research in this domain, Li *et al.* [19] established standard MIRS benchmarks by extending two COVID-19 datasets (QaTa-COV19 [8] and MosMedData+ [25]) with clinician-verified texts, while proposing a hybrid CNN-Transformer framework for feature alignment. Subsequent works have focused on optimizing cross-modal interaction; for instance, LanGuideSeg [30] introduced a simple, hierarchical guidedecoder for progressive visual and textual information fusion, while RecLMIS [14] proposed a conditioned reconstruction model to achieve fine-grained cross-modal alignment.

Despite the substantial performance gains achieved by existing MIRS methods [3,6,7,13,14,19,28] over conventional unimodal models [15,23,31], several critical limitations remain unaddressed. Current architectures predominantly rely on early- or late-fusion strategies that treat the textual input as a monolithic global feature. While this provides a general semantic context, it lacks the granular, token-level interaction required for precise spatial grounding. Although some recent works [28,30] suggest that location-based tokens exert the most influence on performance, a robust model must leverage the synergy between diverse descriptors. For instance, the infection type (*e.g.,* bilateral pulmonary infection) identifies the lesion texture, the quantity (*e.g.,* three infected areas) constrains the range of interest, and the location (*e.g.,* upper-lower left lung) grounds the search space. Furthermore, recent approaches addressing incomplete text via progressive learning [29] assume settings less consistent with routine clinical practice, where standardized reports are usually available. These methods fuse text and image features using standard cross-attention without explicit role differentiation. In addition, the morphological overlap between lesions and healthy tissue also leads to fuzzy boundaries.

To this end, we propose a **R**ole-**C**onditioned **A**lignment (**RCA**) method for MIRS. Our approach centers on a specialized block that refines visual features by exploiting the distinct functional characteristics of clinical text tokens. The process begins with frequency-enhanced upsampling, which restores high-frequency structural details lost during visual feature scaling. These features are

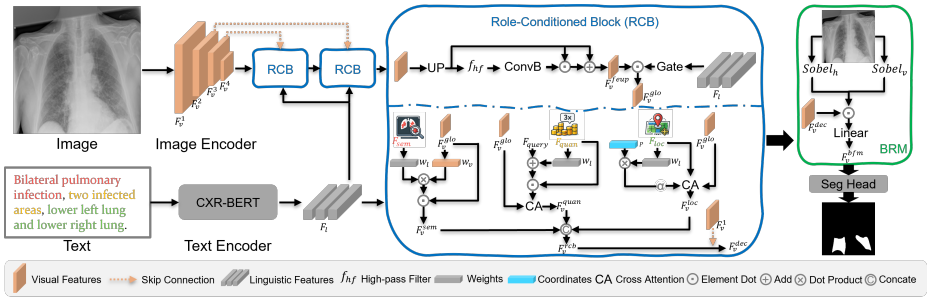


**Fig. 1.** Overall framework of the proposed RCA model. We use pre-trained encoders to extract visual and linguistic features, $F_v$, $F_l$, respectively (Sec. 2.1). The Role-Conditioned Block (RCB in blue area) refines visual features using distinctive roles of text tokens (Sec. 2.2) and the Boundary Refinement Module (BRM in green area) with edge-aware information (Sec. 2.3) further improves segmentation predictions.

then aligned with linguistic features through one global textual context interaction and three parallel, role-conditioned attention mechanisms: a coordinate scoring function for spatial grounding, an energy-based modeling for lesion semantics, and a text-gated lesion query for quantitative priors. To ensure precise delineation, we introduce a boundary refinement module prior to the segmentation head, which leverages edge-aware gradients to disambiguate lesions from surrounding healthy tissue. By aligning specialized textual roles with visual cues, RCA achieves a more granular cross-modal alignment, outperforming previous methods across standard MIRS benchmarks.

Our main contributions are summarized as follows:

- We propose the RCA framework for MIRS, which explicitly models the functional roles of text tokens to facilitate focused cross-modal reasoning.
- We design a boundary refinement module that integrates edge-aware priors to enhance structural delineation in complex morphological scenarios.
- Extensive experiments demonstrate that RCA achieves new state-of-the-art performance across the QaTa-COV19 and MosMedData+ benchmarks.

## 2 Method

To address the challenges of MIRS, we introduce RCA, an architecture designed to explicitly model the functional semantics inherent in clinical referring expressions. Fig. 1 illustrates the overview of the proposed method.

### 2.1 Input Representation

Following previous works [10,27,30], we employ ConvNeXt-Tiny [20] as the visual backbone to extract multi-scale image features. Given an input image $I \in$

$\mathbb{R}^{H\times W\times 3}$, the encoder generates a set of hierarchical feature maps $\{F_v^i\}_{i=1}^4 \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$, where $C_i$ denotes the channel dimension at each respective stage. For the linguistic input, we adopt CXR-BERT [2], a domain-specific encoder pretrained on radiological reports, to extract the textual representation $F_l \in \mathbb{R}^{L\times C_l}$, where $L$ denotes the sequence length and $C_l$ is the feature dimension. Unlike conventional approaches that treat $F_l$ as a monolithic feature, we explicitly partition the text tokens based on their functional roles into three distinct subsets: lesion semantics ($F_{sem} \in \mathbb{R}^{L_1\times C_l}$), quantity ($F_{quan} \in \mathbb{R}^{L_2\times C_l}$), and anatomical location ($F_{loc} \in \mathbb{R}^{L_3\times C_l}$).

### 2.2 Role-Conditioned Block

The role-conditioned block is the core module of our model, which refines cross-modal representations through three sequential stages: frequency-enhanced upsampling (FEUP), role-conditioned interaction (RCI), and feature fusion (FF).

**FEUP.** Standard upsampling operations, such as bilinear interpolation, often blur spatial details and result in over-smoothed boundaries. To recover the lost high-frequency signals caused by upsampling operations, FEUP employs a learnable high-pass filter $f_{hf}$ to extract structural gradients from the upsampled map $F_v^{up}$. We define $f_{hf}$ using a normalized Laplacian [16] kernel (center weight 1; surroundings $-1/8$) to emphasize edges. The filtered output is processed through $3\times3$ convolutions with ReLU and Sigmoid activation ($\sigma$) to generate a frequency compensation map $F_{com}$. This map adaptively modulates the upsampled features via element-wise multiplication ($\odot$), followed by a residual connection and a feed-forward network (FFN) to produce the refined feature $F_v^{feup}$:

$$F_{com} = \sigma(\mathrm{ConvB}(f_{hf}(F_v^{up}))), \tag{1}$$

$$F_v^{feup} = \mathrm{FFN}(F_{com} \odot F_v^{up} + F_v^{up}), \tag{2}$$

**RCI.** RCI performs global linguistic alignment followed by three parallel, role-specific attention mechanisms. Inspired by SENet [12], we first employ a textual gate to dynamically generate channel-wise weights from the complete textual embedding $F_l$, which serve as the linguistic context. This gate, consisting of global average pooling followed by a linear projection ($W_g$) and Sigmoid activation, modulates $F_v^{feup}$ to emphasize contextually relevant channels:

$$F_v^{glo} = \sigma(\mathrm{W}_g\mathrm{AvgPool}(F_l)) \odot F_v^{feup}, \tag{3}$$

For *lesion semantics*, we model the alignment of pathological descriptions as an energy-based distribution. By projecting visual and semantic features into a shared latent space, we compute a normalized energy map that highlights semantically consistent regions:

$$F_v^{sem} = \mathrm{softmax}(\mathrm{W}_l F_{sem} \cdot \mathrm{W}_v F_v^{glo}) \odot F_v^{glo}, \tag{4}$$

where $W_v$ and $W_l$ are learnable projections. This attention allows $F_v^{sem}$ to focus on areas where the energy landscape aligns with the semantic tokens. For *infection quantity*, we introduce a text-gated lesion query mechanism. We define a

set of learnable lesion queries $F_{query} = \{q_i\}_{i=1}^{n}$, where $n$ is the maximum number of potential lesions. The quantity tokens $F_{quan}$ are used to modulate these queries via a text-conditioned shift and a gating signal, effectively adjusting their activation strength based on the clinical description:

$$F'_{query} = (F_{query} + \mathrm{W}_l F_{quan}) \odot \sigma(\mathrm{Proj}(F_{quan})), \tag{5}$$

These quantity-informed queries $F'_{query}$ then interact with the visual features $F_v^{glo}$ through a cross-attention ($CA$) module [24] to generate the quantity-consistent feature map $F_v^{quan}$. For tokens related to *anatomical location*, we propose a location-conditioned coordinate scoring strategy that modulates attention logits with spatial priors, enabling semantically grounded reasoning beyond standard $CA$. We first generate coordinate embeddings $P$ by passing the normalized coordinates at each location $j$ through an MLP($[x_j, y_j]$). A spatial bias is then computed via the dot product of the projected location tokens and the coordinate embeddings. This bias is integrated into the cross-attention mechanism with a learnable scaling factor $\alpha$. To further sharpen positional sensitivity, we apply a self-gated tanh modulation to obtain the location-conditioned features:

$$f_v^{loc} = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} + \alpha \cdot (\mathrm{W}_l F_{loc})P\right)V, \tag{6}$$

$$F_v^{loc} = \tanh(f_v^{loc}) \odot f_v^{loc}, \tag{7}$$

**FF.** Once the role-specific features are obtained, we fuse them to form a unified representation $F_v^{rcb}$. Specifically, the semantic ($F_v^{sem}$), quantitative ($F_v^{quan}$), and spatial ($F_v^{loc}$) cues are integrated via a concatenation-based fusion followed by an aggregation operation Agg($\cdot$). Here, Agg($\cdot$) consists of sequential linear projections, layer normalization, and ReLU activations, designed to map the multimodal responses into a unified feature space. Following [30], $F_v^{rcb}$ are fused with the current-scale visual features to facilitate multi-scale interaction and produce the decoded output $F_v^{dec}$.

### 2.3 Boundary Refinement Module

To improve the delineation between lesions and healthy tissue, we introduce a Boundary Refinement Module (BRM). We first extract normalized edge priors from the input image $I$ using a Sobel [17] operator to compute the gradient magnitude. This prior is fused with the decoded features to enhance boundary contrast. The final refinement is performed by a Mixer consisting of linear layers and ReLU activations:

$$F_v^{edge} = \sigma(\sqrt{(\mathrm{Sobel_h}(I))^2 + (\mathrm{Sobel_v}(I))^2}) \odot F_v^{dec}, \tag{8}$$

$$F_v^{brm} = \mathrm{Mixer}(F_v^{edge}) \odot F_v^{edge}, \tag{9}$$

The output $F_v^{brm}$ is then passed to a segmentation head for the final pixel-level predictions. Following prior works [10,27,30], we adopt Dice loss and Cross-entropy loss as the loss function when training the network.

**Table 1.** Comparison with existing medical image segmentation methods on the QaTa-COV19 [8] and MosMedData+ [25] datasets, reporting performance in %.

| Method | Backbone | Text | QaTa-COV19 | | MosMedData+ | |
|---|---|---|---|---|---|---|
| | | | Dice | mIoU | Dice | mIoU |
| U-Net [23] | CNN | ✗ | 79.02 | 69.46 | 64.60 | 50.73 |
| UNet++ [31] | CNN | ✗ | 79.62 | 70.25 | 71.75 | 58.39 |
| nnUnet [15] | CNN | ✗ | 80.42 | 70.81 | 72.59 | 60.36 |
| TransUNet [5] | Hybrid | ✗ | 78.63 | 69.13 | 71.24 | 58.44 |
| Swin-Unet [4] | Hybrid | ✗ | 78.07 | 68.34 | 63.29 | 50.19 |
| LViT [19] | Hybrid | ✓ | 83.66 | 75.11 | 74.57 | 61.33 |
| LanGuideSeg [30] | CNN | ✓ | 89.78 | 81.45 | - | - |
| LGA [13] | Transformer | ✓ | 84.65 | 76.23 | 75.63 | 62.52 |
| CausalCLIPSeg [7] | Hybrid | ✓ | 85.21 | 76.90 | - | - |
| MAdapter [28] | CNN | ✓ | 90.22 | 82.16 | 78.62 | 64.78 |
| TGCAM [10] | CNN | ✓ | 90.60 | 82.81 | 77.82 | 63.69 |
| MMI-UNet [3] | CNN | ✓ | 90.88 | 83.28 | 78.42 | 64.50 |
| RecLMIS [14] | CNN | ✓ | 85.22 | 77.00 | 77.48 | 65.07 |
| T-S [29] | Hybrid | ✓ | 88.73 | 79.74 | 79.40 | 65.84 |
| NTP-MRISeg [6] | Transformer | ✓ | 91.10 | 83.66 | 79.18 | 65.54 |
| FMISeg [27] | CNN | ✓ | 91.21 | 83.84 | 79.30 | 65.71 |
| RCA (ours) | CNN | ✓ | **92.04** | **84.92** | **80.21** | **67.32** |

## 3 Experiments

### 3.1 Datasets and Evaluation Metric

**Datasets.** The experiments were conducted on two public MIS datasets, QaTa-COV19 [8] and MosMedData+ [25]. Originally restricted to image-mask pairs, these two COVID-19 datasets were transformed into multimodal benchmarks by Li *et al.* [19] through the manual addition of clinician-verified textual descriptions, including the infection's type, quantity, and anatomical location. QaTa-COV19 contains 9,258 chest radiographs (5,716/1,429/2,113 for train/val/test) with pixel-level masks. MosMedData+ comprises 2,729 CT slices (2,183/273/273 for train/val/test) featuring binary masks paired with descriptive texts.
**Evaluation Metric.** Following the standard protocol in previous MIRS studies, we report the Dice coefficient (Dice) and mean intersection-over-union (mIoU) scores to evaluate model performance across diverse experimental configurations.

### 3.2 Implementation Details

We used ConvNeXt-Tiny [20] as the image encoder. To assign text tokens to their respective roles, we employed a rule-based parsing strategy that categorizes tokens based on the clinical report structure, ensuring consistent mapping into the semantic, quantity, and location modules. To capture domain-specific semantic information, we utilized CXR-BERT [2], to generate textual embeddings. Consistent with the experimental protocols of previous studies, the input

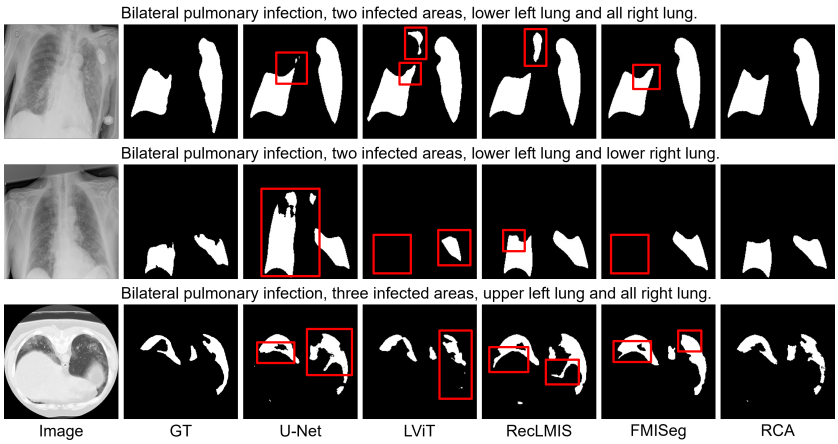


**Fig. 2.** Qualitative comparison of different methods (U-Net [23], LViT [19], RecLMIS [14], FMISeg [27] and our proposed RCA) on the QaTa-COV19 [8] (row 1 - 2) and MosMedData+ [25] (row 3) datasets. GT represents the ground truth.

image resolution was set to $224 \times 224$. All models were trained using the AdamW [21] optimizer with an initial learning rate of $3e$-4, which is reduced to $1e$-6. The batch size and number of training epochs were set to 32 and 100, respectively. All experiments were conducted using PyTorch on a single NVIDIA H100 GPU.

### 3.3 Comparison with State-of-the-art Methods

**Compared Methods.** To verify the effectiveness of the proposed method, we compared it with previous state-of-the-art *unimodal* and *multimodal* medical image segmentation approaches. For *unimodal* methods, we evaluated U-Net [23], UNet++ [31], nnUnet [15], TransUnet [5], and Swin-Unet [4]. For *multimodal* methods, we compared against a comprehensive set of MIRS methods, including LViT [19], LanGuideSeg [30], LGA [13], CausalCLIPSeg [7], MAdapter [28], TGCAM [10], RecLMIS [14], T-S [29], NTP-MRISeg [6], and FMISeg [27].

**Quantitative Results.** The quantitative results are reported in Table 1. Our proposed RCA achieved state-of-the-art results across both benchmarks, outperforming the previous leading method, FMISeg, by 1.08% and 1.61% in mIoU, respectively. On the QaTa-COV19 dataset, compared with *unimodal* methods (*e.g.,* U-Net and nnUnet), RCA showed substantial performance gains (exceeding 10%), underscoring the critical value of linguistic guidance in medical segmentation. Notably, RCA also surpassed recent *multimodal* methods that utilize complex cross-modal interactions. For example, our method achieved a 92.04% Dice score, significantly outperforming LViT (83.66%) and RecLMIS (85.22%). The performance margin was particularly evident on the MosMedData+ dataset, where RCA achieved 80.21% Dice and 67.32% mIoU, representing a gain of over 5% compared to LViT. In addition, RCA obtained a clear lead (0.91% Dice and

**Table 2.** Ablation study on the main modules and backbones.

| Ablation | Model | QaTa-COV19 | | MosMedData+ | |
|---|---|---|---|---|---|
| | | Dice | mIoU | Dice | mIoU |
| # 1 | Baseline | 85.87 | 76.15 | 76.34 | 63.06 |
| # 2 | # 1 + FEUP | 86.05 | 76.44 | 76.52 | 63.40 |
| # 3 | # 2 + semantic role | 88.73 | 79.59 | 78.17 | 65.02 |
| # 4 | # 2 + quantity role | 89.16 | 80.07 | 78.46 | 65.28 |
| # 5 | # 2 + location role | 90.55 | 81.19 | 78.93 | 65.70 |
| # 6 | # 2 + all roles + FF | 91.84 | 84.42 | 79.85 | 66.87 |
| # 7 | # 6 + BRM (ours) | **92.04** | **84.92** | **80.21** | **67.32** |
| # 8 | ResNet [11] + CLIP [22] | 91.36 | 83.25 | 79.37 | 66.54 |
| # 9 | U-Net [23] + BERT [9] | 91.80 | 84.63 | 79.84 | 67.03 |
| # 10 | ConvNeXt [20] + CXR-BERT [2] | **92.04** | **84.92** | **80.21** | **67.32** |

1.61% mIoU) over FMISeg. These results validate our role-conditioned design, demonstrating its efficacy in leveraging textual descriptions to refine lesion prediction and resolve spatial ambiguities across both X-ray and CT modalities.

**Qualitative Results.** As illustrated in Fig. 2, RCA yielded segmentation with significantly fewer false positives and more precise boundaries than competing methods (highlighted in red bounding boxes). By explicitly modeling linguistic roles, our framework effectively disambiguated diseased regions from overlapping healthy tissue. These visual improvements confirm that our role-conditioned design successfully translates clinical descriptions into precise guidance for MIS.

### 3.4 Ablation Studies

**Ablation Study on Main Modules.** We evaluated the contribution of each component within RCA as reported in the upper section of Table 2. Using a standard cross-attention-based model as a baseline, we observe that FEUP (#2) provided a steady performance gain by restoring high-frequency structural details. The introduction of individual role-conditioned modules (#3–#5) consistently improved results across both benchmarks. Specifically, the location-conditioned coordinate scoring (#5) provided the most significant boost, increasing the Dice score by 4.5% on QaTa-COV19. This underlines the critical role of spatial grounding in disambiguating diseased regions, which is also verified by prior works [28,30]. Integrating all three roles with FF (#6) further elevated performance, demonstrating that semantic, quantitative, and spatial cues offer complementary information. Finally, the inclusion of BRM (#7) achieved the best performance by effectively refining lesion boundaries.

**Ablation Study on Backbones.** The lower section of Table 2 evaluates RCA across various backbone combinations. Our framework consistently achieved high performance regardless of the specific encoders used, demonstrating that the role-conditioned design successfully generalizes across different backbones. Notably, the ConvNeXt + CXR-BERT configuration (#10) yielded the best results.

## 4 Conclusion

In this paper, we introduce a novel method RCA for medical image referring segmentation. By explicitly partitioning textual descriptions into semantic, quantitative, and spatial roles, our method facilitates a more granular cross-modal alignment compared to monolithic embedding approaches. The proposed boundary refinement module ensures that fine-grained structural details help accurate delineation. Experiments on the QaTa-COV19 and MosMedData+ benchmarks demonstrate that RCA achieves state-of-the-art performance, proving the effectiveness of role-conditioned alignment in complex medical scenarios.

**Acknowledgments.** This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK Research and Innovation (UKRI), under Grant EP/X01441X/1.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European Conference on Computer Vision. pp. 1–21 (2022)
3. Bui, P.N., Le, D.T., Choo, H.: Visual-textual matching attention for lesion segmentation in chest images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 702–711 (2024)
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. pp. 205–218 (2022)
5. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
6. Chen, X., Wang, Y., Pang, G., Hao, J., Yue, C., Zhou, L., Li, Y.: Medical referring image segmentation via next-token mask prediction. arXiv preprint arXiv:2511.05044 (2025)
7. Chen, Y., Wei, M., Zheng, Z., Hu, J., Shi, Y., Xiong, S., Zhu, X.X., Mou, L.: Causalclipseg: Unlocking clip's potential in referring medical image segmentation with causal intervention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 77–87 (2024)
8. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: IEEE International Conference on Image Processing. pp. 2306–2310 (2022)

9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the conference of the North American chapter of the Association for Computational Linguistics. pp. 4171–4186 (2019)
10. Guo, Y., Zeng, X., Zeng, P., Fei, Y., Wen, L., Zhou, J., Wang, Y.: Common vision-language attention for text-guided medical image segmentation of pneumonia. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 192–201 (2024)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
13. Hu, J., Li, Y., Sun, H., Song, Y., Zhang, C., Lin, L., Chen, Y.W.: 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. pp. 610–620 (2024)
14. Huang, X., Li, H., Cao, M., Chen, L., You, C., An, D.: Cross-modal conditioned reconstruction for language-guided medical image segmentation. IEEE Transactions on Medical Imaging (2024)
15. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
16. Jain, R., Kasturi, R., Schunck, B.G., et al.: Machine vision, vol. 5. McGraw-hill New York (1995)
17. Kanopoulos, N., Vasanthavada, N., Baker, R.L.: Design of an image edge detection filter using the sobel operator. IEEE Journal of Solid-state Circuits **23**(2), 358–367 (1988)
18. Li, K., Vosselman, G., Yang, M.Y.: Scale-wise bidirectional alignment network for referring remote sensing image segmentation. ISPRS Journal of Photogrammetry and Remote Sensing **226**, 350–363 (2025)
19. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging **43**(1), 96–107 (2023)
20. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
21. Loshchilov, I., Hutter, F., et al.: Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101 (2017)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241 (2015)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)

25. Vladzymyrskyy, A., Gonchar, A., Chernina, V.Y.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465 (2020)
26. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022)
27. Yu, B., Yang, J., Du, Z., Huang, Y., Li, C., Wang, L.: Frequency-domain multi-modal fusion for language-guided medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 278–288 (2025)
28. Zhang, X., Ni, B., Yang, Y., Zhang, L.: Madapter: A better interaction between image and language for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 425–434 (2024)
29. Zhao, P., Hou, Y., Yan, Z., Huo, S.: Text-driven medical image segmentation with text-free inference via knowledge distillation. IEEE Transactions on Instrumentation and Measurement (2025)
30. Zhong, Y., Xu, M., Liang, K., Chen, K., Wu, M.: Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 724–733 (2023)
31. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International Workshop on Deep Learning in Medical Image Analysis. pp. 3–11 (2018)