



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

HiProS: A Hierarchical Prompt Selection for Stage-wise Vision–Language Alignment in Medical Image Segmentation

Da-Hee Kim, Jongbin Chae, Seungyeon Lee, and Dong-Gyu Lee*

Department of Artificial Intelligence,
Kyungpook National University, Daegu, Republic of Korea
{heesdajoy,kalibianw,statai3237,dglee}@knu.ac.kr

Abstract. Recent vision–language models for medical image segmentation leverage radiology reports as auxiliary supervision. However, existing methods inject text cues uniformly across decoder stages, overlooking the hierarchical semantic demands of different visual abstraction levels. This limits precise alignment between global context and fine-grained lesion representations, particularly for diffuse abnormalities. To address this issue, we propose Hierarchical Prompt Selector (HiProS), a vision–language segmentation framework that performs stage-conditioned textual token selection for local–global alignment. HiProS introduces a Stage-wise Prompt Selector that explicitly selects and suppresses textual tokens according to the abstraction level of each decoder stage, rather than propagating all text cues through uniform attention. The selected tokens are then integrated with multi-scale visual features via a Fusion Alignment Module, enabling structured cross-modal interaction at each decoding stage. Importantly, HiProS operates with automatically generated textual cues, avoiding reliance on manually curated clinical reports and enabling scalability where expert annotations are unavailable. Experiments on QaTa-COV19 and MosMedData+ demonstrate consistent improvements over strong vision–language baselines, with competitive results. Notably, HiProS shows pronounced gains in segmenting diffuse lesions on MosMedData+, indicating that stage-conditioned textual selection mitigates semantic misalignment in fine-grained segmentation.

Keywords: Medical Image Segmentation · Vision–Language Model · Multi-modal Fusion

1 Introduction

Medical image segmentation focuses on accurately identifying regions of interest, such as tumors or inflammatory lesions, and plays a crucial role in computer-aided diagnosis and treatment planning [6, 26]. With the advancement of deep learning, several vision-based architectures have been explored for medical image segmentation [27, 25]. However, these existing approaches [31, 4, 5] remain limited

* Corresponding author.

in capturing complex clinical semantics, as they primarily rely on pixel-level visual supervision without diagnostic textual context. Consequently, the lack of semantic grounding limits generalization in real-world clinical settings.

To mitigate insufficient semantic grounding, recent studies have incorporated textual information to guide medical image segmentation. Early text-guided approaches [24, 12, 17, 11] integrate textual representations into visual encoders or decoders to enhance cross-modal alignment. C2FVL [24] injects textual cues via a coarse-to-fine attention strategy, while LViT [17] models multi-scale vision–language interactions using a hybrid CNN–Transformer architecture. Subsequently, vision–language segmentation approaches [10, 23, 30, 13] introduced dynamic interaction mechanisms that iteratively refine textual representations conditioned on visual feedback. TGCAM [10] progressively fuses visual and textual features through iterative cross-attention, whereas RecLMIS [13] updates text–visual representations via recurrent cross-modal interaction.

While prior radiology studies [9, 15] have emphasized the importance of structured clinical reports, existing segmentation methods fail to explicitly model the hierarchical structure of medical reports. As illustrated in Fig. 1, *Findings* describe lesion locations and fine-grained morphological details, whereas *Impressions* summarize high-level diagnostic interpretations. These components correspond to the progressive abstraction of the decoding hierarchy, where early stages emphasize detailed morphology and later stages capture global context. Despite this correspondence, prior approaches typically share an identical textual representation across all decoding stages or repeatedly reweight it, without explicit stage-conditioned semantic allocation. As a result, local and global semantics fail to be disentangled, limiting precise alignment with multi-scale visual representations. While IMITATE [18] exploits hierarchical report semantics to align multi-level visual representations during pre-training, explicitly allocating such semantics within the segmentation decoder remains underexplored.

To address these limitations, we propose Hierarchical Prompt Selector (HiProS), a vision–language segmentation framework that aligns clinical report semantics with stage-specific visual abstraction. HiProS associates fine-grained observational cues with early decoding stages and high-level diagnostic interpretations with later stages, reflecting the complementary roles of *Findings* and *Impressions*. Specifically, we introduce a Stage-wise Prompt Selector (SPS) that performs explicit token-level selection and suppression at each decoding stage, ensuring that semantically relevant textual cues guide the appropriate level of visual representation. The selected tokens are integrated with multi-scale visual features via a Fusion Alignment Module (FAM) to enable structured cross-modal interaction. In addition, HiProS leverages a Vision–Language Model (VLM) to automatically generate structured medical reports, reducing reliance on clinician-authored annotations and establishing a scalable supervision paradigm.

Our contributions are summarized as follows: (1) We propose HiProS, a stage-aware vision–language segmentation framework that structurally aligns hierarchical textual semantics with progressive visual abstraction in the decoding hierarchy. (2) We introduce a Stage-wise Prompt Selector (SPS) that performs

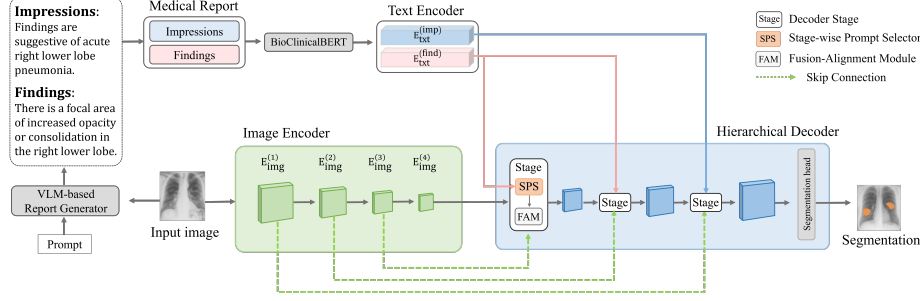


Fig. 1. The overall architecture of the proposed HiProS.

explicit token-level selection and suppression, allocating semantically relevant textual cues to each decoding stage. (3) We establish a scalable supervision framework using VLM-generated structured *Findings* and *Impressions*, reducing reliance on clinician-authored reports. (4) Experiments on QaTa-COV19 and MosMedData+ demonstrate consistent state-of-the-art performance.

2 Methodology

Fig. 1 illustrates the overall architecture of HiProS. Given an input medical image, a VLM-based report generator produces a hierarchical report composed of *Findings* and *Impressions*. The generated text is tokenized and encoded into token embeddings, while the image encoder extracts multi-level visual features. At each decoding stage, text embeddings corresponding to the visual abstraction level are allocated in a stage-conditioned manner, and the SPS performs token-level selection by suppressing semantically irrelevant tokens. The selected token embeddings are aligned with visual features via the FAM, enabling stage-wise local-global semantic integration. The fused multimodal representations are progressively refined to predict the segmentation mask.

2.1 Text and Image Encoder

We model hierarchical clinical semantics using MiniGPT-Med [1] to generate structured radiology reports consisting of *Findings* for detailed anatomical observations and *Impressions* for high-level diagnostic interpretations. Report generation uses only the input image and a fixed prompt, without access to ground-truth masks, lesion labels, or dataset annotations, thus preventing information leakage from segmentation supervision. The generated reports are encoded using a pretrained BioClinicalBERT [2] model to produce token embeddings $E_{\text{txt}}^{\text{find}} \in \mathbb{R}^{B \times M \times d}$ and $E_{\text{txt}}^{\text{imp}} \in \mathbb{R}^{B \times M \times d}$ for *Findings* and *Impressions*, respectively, where B is the batch size, M is the sequence length, and d is the embedding dimension. For visual encoding, we use ConvNeXt-Tiny [19] to extract hierarchical visual representations from the input medical image, producing multi-scale

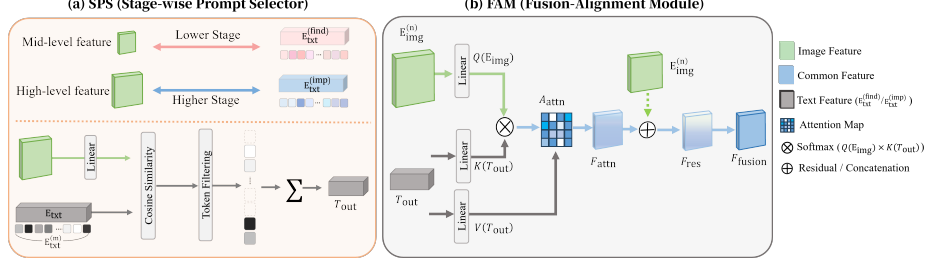


Fig. 2. Details of the proposed modules. (a) SPS performs stage-conditioned token selection. (b) FAM fuses selected tokens and visual features via cross-attention.

feature maps $E_{\text{img}} = \{E_{\text{img}}^{(1)}, E_{\text{img}}^{(2)}, \dots, E_{\text{img}}^{(N)}\} \in \mathbb{R}^{B \times C \times H \times W}$, where N denotes the number of scales, C is the number of channels, and H and W denote the spatial height and width, respectively. Each feature map corresponds to a different spatial resolution and semantic abstraction level, providing multi-scale visual context for subsequent stage-wise decoding.

2.2 Stage-wise Prompt Selector

As illustrated in Fig. 2(a), the SPS performs stage-conditioned token selection to align textual cues with the visual abstraction level of each decoder stage. Lower decoder stages utilize $E_{\text{txt}}^{\text{find}}$, which reflects localized morphological information, whereas higher stages leverage $E_{\text{txt}}^{\text{imp}}$, capturing global lesion context. The tokens assigned to each stage are subsequently subjected to stage-wise selection.

For decoder stage n , the visual embedding is given as $E_{\text{img}}^n \in \mathbb{R}^{B \times C_n \times H_n \times W_n}$. It is summarized using Global Average Pooling (GAP) and projected into a shared embedding space:

$$z_{\text{img}}^n = \text{GAP}(E_{\text{img}}^n)W_n, \quad W_n \in \mathbb{R}^{C_n \times d}, \quad (1)$$

where d denotes the shared embedding dimension. The semantic relevance between the projected visual representation $z_{\text{img}}^n \in \mathbb{R}^{B \times d}$ and each textual token embedding $E_{\text{txt}}^m \in \mathbb{R}^{B \times d}$, where m indexes the textual tokens, is computed using cosine similarity:

$$S_{n,m} = \frac{z_{\text{img}}^n \cdot E_{\text{txt}}^m}{\|z_{\text{img}}^n\| \|E_{\text{txt}}^m\|}. \quad (2)$$

Based on these similarity scores $S_{n,m}$, SPS applies hard gating using a learnable stage-specific threshold τ_n :

$$G_{n,m} = \mathbb{I}(S_{n,m} \geq \tau_n), \quad (3)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Tokens whose similarity exceeds τ_n are selected while the others are suppressed, enabling explicit filtering. Each stage maintains a learnable threshold to allow adaptive stage-wise token selection.

Since the indicator function is non-differentiable, the backward pass employs a sigmoid-based surrogate gradient via a Straight-Through Estimator [3, 28]:

$$\frac{\partial G_{n,m}}{\partial \tau_n} \approx -\sigma(S_{n,m} - \tau_n) \left(1 - \sigma(S_{n,m} - \tau_n)\right), \quad (4)$$

where $\sigma(\cdot)$ denotes the sigmoid function. This surrogate formulation enables gradient-based optimization of τ_n in Eq. 4. The threshold τ_n is initialized to 0.01 to encourage inclusive token selection during the early stage of training and is progressively adapted thereafter. Finally, the selected tokens are aggregated via similarity-normalized weighted pooling to produce a compact stage-specific representation T_{out}^n , which guides subsequent multimodal fusion.

2.3 Fusion-Alignment Strategy

The SPS selects stage-relevant textual semantics, but does not explicitly model their spatial correspondence with visual features. As illustrated in Fig. 2(b), we introduce the FAM that aligns the stage-specific textual cues selected by the SPS with visual features, progressively integrating both local and global textual semantics. At each stage, the FAM takes the visual embedding E_{img} and the stage-specific representation T_{out} generated by the SPS as inputs. The visual features are projected as queries Q , while the textual representations are linearly projected as keys K and values V . Cross-attention then injects textual guidance into the visual feature space:

$$A_{\text{attn}} = \text{softmax} \left(\frac{Q(E_{\text{img}}) \cdot K(T_{\text{out}})^\top}{\sqrt{d}} \right), \quad F_{\text{attn}} = A_{\text{attn}} \cdot V(T_{\text{out}}), \quad (5)$$

where A_{attn} denotes the attention weights between spatial visual features and stage-specific textual tokens, and d denotes the attention dimension. Unlike conventional cross-attention that shares a single textual representation across stages, the FAM enables hierarchy-consistent cross-modal alignment via SPS-selected tokens. The aligned feature F_{attn} is concatenated with the original visual embedding along the channel dimension to form a residual multimodal representation F_{res} . The fused multimodal representation is then produced as

$$F_{\text{fusion}} = \text{Conv}(F_{\text{res}}), \quad (6)$$

where Conv denotes a refinement block consisting of a 3×3 convolution, batch normalization, and ReLU activation. Finally, F_{fusion} is passed to the next decoder stage, and this alignment-fusion process is repeated hierarchically across stages.

2.4 Training Objective

The total training objective is formulated as a weighted combination of Dice loss, cross-entropy loss, and boundary-sensitive loss [8]:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Dice}} + \lambda_2 \mathcal{L}_{\text{CE}} + \lambda_3 \mathcal{L}_{\text{Boundary}}. \quad (7)$$

$\mathcal{L}_{\text{Dice}}$ and $\mathcal{L}_{\text{CE}}$ provide complementary region-level supervision, while $\mathcal{L}_{\text{Boundary}}$ enhances contour precision by emphasizing lesion boundary consistency.

Table 1. Quantitative comparison of segmentation results. “Hybrid” indicates CNN–Transformer integration. † denotes HiProS using LViT single-text reports, and ‡ denotes HiProS using LLaVA-Med [16] generated hierarchical reports. The best results are shown in **bold**, and the second-best results are underlined.

Method	Backbone Text		Text Strategy	QaTa-COV19		MosMedData+	
				Dice(%)	mIoU(%)	Dice(%)	mIoU(%)
U-Net++ [31]	CNN	×	-	79.62	70.25	71.75	58.39
nnUNet [14]	CNN	×	-	80.42	70.81	72.59	60.36
TransUNet [5]	Hybrid	×	-	78.63	69.13	71.24	58.44
C2FVL [24]	CNN	✓	Raw	83.40	74.62	74.56	61.15
LViT [17]	Hybrid	✓	Raw	83.66	75.11	74.57	61.33
LGA [11]	SAM	✓	Raw	84.65	76.23	75.63	62.52
FMISeg [29]	CNN	✓	Raw	<u>91.21</u>	<u>83.84</u>	<u>79.30</u>	<u>65.71</u>
RecLMIS [13]	CNN	✓	Interactive	85.22	77.00	77.48	65.07
TGCAM [10]	CNN	✓	Interactive	90.60	82.81	77.82	63.69
MAdapter [30]	CNN	✓	Interactive	90.22	82.16	78.62	64.78
HiProS (Ours)†	CNN	✓	w/o hierarchical text	92.12	86.85	80.25	72.18
HiProS (Ours)‡	CNN	✓	Hierarchical + Selective	92.42	86.77	82.15	74.29
HiProS (Ours)	CNN	✓	Hierarchical + Selective	92.64	87.10	84.27	76.50

3 Experiment

3.1 Experimental Setup

Datasets. We conducted experiments on two public medical image segmentation datasets: QaTa-COV19 [7] and MosMedData+ [21]. QaTa-COV19 contains 9,258 chest X-ray images with pixel-level infection masks, split into 5,716 training, 1,429 validation, and 2,113 testing samples following the LViT protocol [17]. MosMedData+ consists of 2,729 axial CT slices with infection annotations, divided into 2,183 training, 273 validation, and 273 testing samples using the same protocol. For text supervision, we use the medical reports from [17] as single-text inputs and generate structured reports using MiniGPT-Med [1], with additional reports generated by LLaVA-Med [16] to assess cross-VLM robustness.

Implementation Details. We implemented HiProS in PyTorch [22] with an NVIDIA RTX A6000 GPU (48 GB). The model was optimized with AdamW [20] using an initial learning rate of 1×10^{-4} and a cosine annealing schedule for 50 epochs, with a batch size of 32. Input images were resized to 224×224 and normalized using ImageNet statistics. The decoder comprises four hierarchical stages. *Impressions* are aligned with the highest abstraction stage, while *Findings* are assigned to the remaining lower stages according to their semantic granularity. The loss hyperparameters λ_1 , λ_2 , and λ_3 were set to 1.0, 1.0, and 0.1, respectively. We adopt mean Intersection over Union (mIoU) and Dice coefficient for evaluation. The resulting HiProS model contains 48.4M parameters and requires 13.9 GFLOPs, with an inference time of 7.27 ms per image.

Fig. 3. Qualitative visualization of segmentation results on QaTa-COV19 and MosMed-Data+, including chest X-ray and CT images.

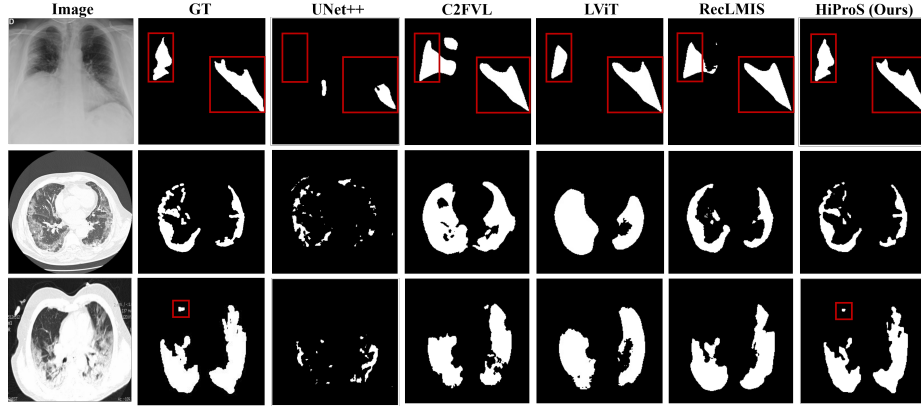
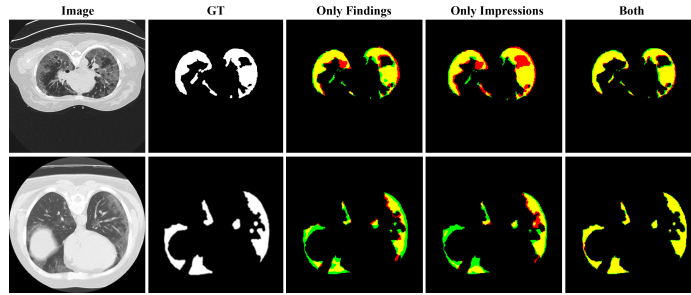


Fig. 4. Qualitative results on the MosMedData+ dataset. Yellow, Green, and Red denote true positive, false negative, and false positive regions, respectively.



3.2 Comparison with State-of-the-art Methods

We compared HiProS with recent segmentation approaches, including vision-only (U-Net++ [31], nnUNet [14], TransUNet [5]), text-assisted (C2FVL [24], LViT [17], FMISeg [29]), and text-interactive models (RecLMIS [13], TGCAM [10], MAdapter [30]). As shown in Table 1, HiProS achieves state-of-the-art performance on both datasets. Compared with TGCAM, MAdapter, and FMISeg, which share the same ConvNeXt-Tiny backbone, HiProS yields improvements. Specifically, on MosMedData+, it surpasses FMISeg by 4.97% in Dice and 10.79% in mIoU. These results highlight the effectiveness of the stage-wise text selection mechanism. Moreover, HiProS[†], using the same text input as prior text-assisted methods, outperforms previous baselines, while the full model further improves performance through hierarchical stage-conditioned text modeling. HiProS[‡] maintains stable performance across reports generated by different medical VLMs, indicating robustness to the report generator. As illustrated in Fig. 3, HiProS

Table 2. Ablation of textual components on QaTa-COV19 and MosMedData+.

Text Parts	QaTa-COV19		MosMedData+	
	Dice(%)	mIoU(%)	Dice(%)	mIoU(%)
No Text	81.82	71.88	70.39	64.05
Only Findings	92.44	86.79	81.53	73.89
Only Impressions	92.13	86.53	80.70	73.47
Full Text	92.48	86.85	80.37	72.61
Swap	91.82	86.28	76.74	69.81
Findings & Impressions	92.64	87.10	84.27	76.50

Table 3. Effect of SPS and FAM on MosMedData+.

Backbone	SPS	FAM	MosMedData+	
			Dice(%)	mIoU(%)
✓			70.39	64.05
✓	✓		79.46	71.93
✓	✓	✓	84.27	76.50

Table 4. Ablation on FAM design on MosMedData+.

Ablation	Dice(%)	mIoU(%)
HiProS (w/o FAM)	79.46	71.93
Simple Concat	80.73	72.83
FAM (w/o Skip)	82.11	74.68
FAM (w/o CA)	81.48	73.95
FAM (Ours)	84.27	76.50

produces stable and precise segmentation results, especially on MosMedData+, where lesions are small or diffuse. In contrast, C2FVL and RecLMIS tend to misclassify normal regions as lesions or miss subtle infection areas.

3.3 Ablation Study

We conducted ablation studies to evaluate each component. The text configuration results (Table 2) show that hierarchically separating *Findings* and *Impressions* across decoder stages consistently outperforms using a single text representation. In contrast, cross-assigning the semantic hierarchy degrades performance, particularly on MosMedData+. These results confirm the effectiveness of stage-conditioned text utilization and the importance of aligning textual hierarchy with decoder abstraction levels. At the token level, SPS tends to select localized anatomical terms such as “right lower lobe consolidation” and “perihilar region” at lower stages, while selecting diagnostic summary terms such as “pneumonia” and “moderate severity” at higher stages. Qualitative results in Fig. 4 indicate that hierarchical textual supervision reduces both local false negatives and global over-segmentation. Module-level ablation results (Table 3) show that SPS and FAM improve Dice by 9.1% and 13.9%, respectively, over the backbone-only model, demonstrating their complementary contributions to stage-wise text selection and cross-modal fusion. Furthermore, the FAM analysis (Table 4) shows that integrating skip connections (Skip) and cross-attention (CA) over SPS-selected stage-specific tokens is more effective than simple feature concatenation, supporting stage-conditioned multi-scale cross-modal interaction.

4 Conclusion

In this paper, we propose HiProS, a hierarchical text utilization strategy for medical image segmentation. Unlike conventional approaches that inject a single textual representation across decoding stages, HiProS aligns the hierarchical semantics of *Findings* and *Impressions* with the abstraction level of each decoder stage, thereby integrating global and local semantics. As a result, HiProS achieves consistent performance improvements on both datasets, with particularly notable gains under challenging conditions such as ambiguous boundaries or small lesion sizes. Future work will focus on refining hierarchical text representations and extending the framework to 3D medical image segmentation.

Acknowledgments. This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-Information Technology Research Center (ITRC) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2024-00437718), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2025-02214941).

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

References

1. Alkhaldi, A., Alnajim, R., Alabdullatef, L., Alyahya, R., Chen, J., Zhu, D., Alsinan, A., Elhoseiny, M.: Minigpt-med: Large language model as a general interface for radiology diagnosis. arXiv preprint arXiv:2407.04106 (2024)
2. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. In: Proceedings of the 2nd clinical natural language processing workshop. pp. 72–78 (2019)
3. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432 (2013)
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. Springer (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, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. Medical image analysis **79**, 102444 (2022)
7. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2306–2310. IEEE (2022)
8. Du, J., Guan, K., Liu, P., Li, Y., Wang, T.: Boundary-sensitive loss function with location constraint for hard region segmentation. IEEE Journal of Biomedical and Health Informatics **27**(2), 992–1003 (2022)

9. Gershanik, E.F., Lacson, R., Khorasani, R.: Critical finding capture in the impression section of radiology reports. In: AMIA Annual Symposium Proceedings. vol. 2011, p. 465 (2011)
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. Springer (2024)
11. 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. Springer (2024)
12. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3942–3951 (2021)
13. 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)
14. 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)
15. Lee, K., Yoon, S., Lim, H.: Clarifid: Improving radiology report generation by reinforcing clinically accurate impressions and enforcing detailed findings. *Expert Systems with Applications* p. 130633 (2025)
16. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* **36**, 28541–28564 (2023)
17. 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)
18. Liu, C., Cheng, S., Shi, M., Shah, A., Bai, W., Arcucci, R.: Imitate: Clinical prior guided hierarchical vision-language pre-training. *IEEE Transactions on Medical Imaging* **44**(1), 519–529 (2024)
19. 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)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017)
21. Morozov, S.P., Andreychenko, A.E., Pavlov, N.A., Vladzimirskyy, A., Ledikhova, N.V., Gomboleviskiy, V.A., Blokhin, I.A., Gelezhe, P.B., Gonchar, A., Chernina, V.Y.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465* (2020)
22. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019)
23. Rahman, M.M., Rahman, S., Bhatt, S., Faezipour, M.: Text-assisted vision model for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics* (2025)

24. Shan, D., Li, Z., Chen, W., Li, Q., Tian, J., Hong, Q.: Coarse-to-fine covid-19 segmentation via vision-language alignment. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
25. Tang, C., Zeng, X., Zhou, L., Zhou, Q., Wang, P., Wu, X., Ren, H., Zhou, J., Wang, Y.: Semi-supervised medical image segmentation via hard positives oriented contrastive learning. *Pattern Recognition* **146**, 110020 (2024)
26. Tsuneki, M.: Deep learning models in medical image analysis. *Journal of Oral Biosciences* **64**(3), 312–320 (2022)
27. Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., Wang, Y.: Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Analysis* **79**, 102447 (2022)
28. Yin, P., Lyu, J., Zhang, S., Osher, S., Qi, Y., Xin, J.: Understanding straight-through estimator in training activation quantized neural nets. *arXiv preprint arXiv:1903.05662* (2019)
29. 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. Springer (2025)
30. 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. Springer (2024)
31. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging* **39**(6), 1856–1867 (2019)