Text as the Sieve: Information Bottleneck Guided Prototype Sieving for Text-Guided Medical Image Segmentation

Jiake Long[1], Xinyi Zeng[1], and Yan Wang[1(✉)]

School of Computer Science, Sichuan University, Chengdu, China
wangyanscu@hotmail.com

Abstract. Accurate pulmonary lesion segmentation is vital for clinical diagnosis, yet conventional single-modal models often struggle to leverage the rich semantic priors available in diagnostic reports. While multimodal approaches have emerged to bridge this gap, recent research has transitioned from dense additive fusion to redundancy-reduction strategies such as token pruning. However, existing pruning-based methods remain suboptimal due to a mismatch in Token Scale—pruning concise text instead of redundant visual tokens—and a Delayed Interaction paradigm that allows background noise to propagate through the encoder. To address these limitations, we propose Text-driven Semantic Information-sieve (TeSi), an Information Bottleneck (IB)-inspired framework that shifts the focus to early encoder-side sieving. Specifically, to resolve the scale mismatch, an IB-Sieve mechanism treats text as a semantic filter to rank and hard-select informative visual prototypes; and to reform the interaction paradigm, our framework performs early visual sieving in the encoder, followed by an asymmetric decoder with adaptive context fusion that selectively refines textual guidance for precise boundary recovery without re-introducing noise. Extensive experiments on the QaTa-COV19-v2 and MosMedData+ datasets demonstrate that TeSi achieves competitive performance compared to current state-of-the-arts. Moreover, our method provides improved clinical interpretability through explicit prototype-to-report alignment, validating the effectiveness of the proposed sieving paradigm.

Keywords: Medical Image Segmentation · Information Bottleneck · Multimodal Learning · Prototype Learning.

1 Introduction

Pulmonary diseases represent a significant global health challenge, necessitating accurate lesion segmentation for diagnosis [1]. While deep learning models (CNNs [2–5], Transformers [6–8]) have achieved success, they often struggle with ambiguous boundaries from visual data alone. Clinically, images are paired with professional reports (e.g., "infection in the lower left lung") containing rich semantic priors. Integrating textual guidance is thus a promising avenue to enhance segmentation [9, 10].

Early multimodal methods [9–11] typically adopt an "additive" paradigm, utilizing dense cross-attention to mix linguistic and visual features. Despite their effectiveness, they overlook the redundancy mismatch: diagnostic texts are concise and localized, while medical images include vast irrelevant background [12]. Treating all tokens equally incurs computational waste and noise. To mitigate this, both general computer vision and medical imaging have shifted towards *Token Pruning* [13] and *Information Sieving* to enhance fusion efficiency. For instance, Grounding DINO [14] demonstrates the efficacy of selecting relevant tokens for efficient inference in natural images. This design paradigm has recently been migrated to the medical domain: NMSW [15] introduces differentiable sampling to discard redundant 3D patches for efficient volumetric segmentation, while ABP [16] pioneers token pruning to remove irrelevant textual information. Compared to dense interaction, these approaches share a common advantage: they significantly reduce computational overhead and enhance fusion specificity.

However, applying existing pruning strategies to text-guided medical segmentation remains suboptimal due to two critical mismatches regarding *Token Scale* and *Interaction Paradigm*. First, regarding *Token Scale*, methods like ABP primarily focus on pruning textual tokens. In reality, diagnostic reports are inherently concise; the massive redundancy actually lies in the *visual* modality (e.g., irrelevant background areas). Filtering text while leaving high-resolution visual backgrounds unfiltered fails to target the root source of noise. Second, regarding the *Interaction Paradigm*, most existing methods adopt a "late fusion" strategy, introducing textual guidance only in the decoding stage. This implies that the encoder operates blindly, processing massive background noise alongside lesion features. Consequently, redundancy propagates through deep layers, entangling with semantic features and corrupting the representation before the late-stage textual guidance can effectively intervene, rendering subsequent cross-modal refinements merely "remedial" rather than preventive.

To address these limitations, we propose Text-driven Semantic Information-sieve (TeSi), an Information Bottleneck (IB)-inspired framework [17] that improves multimodal interaction through early visual redundancy suppression and stable semantic reconstruction. Our contributions are threefold: (i) To resolve the *Token Scale* mismatch, we introduce an IB-Sieve mechanism that operates on visual prototypes. Instead of pruning text, we use text embeddings to rank and hard-select informative visual prototypes, explicitly discarding background noise into a residual token. (ii) To resolve the *Interaction Paradigm* issue, we establish a proactive interaction flow comprising *early* encoder-side sieving and an Asymmetric Decoder. By suppressing background noise at the source, we encourage the bottleneck representation to retain task-relevant lesion semantics. Unlike symmetric designs that complicate the visual stream, our decoder selectively prunes textual tokens based on visual states to guide boundary refinement, preventing semantic drift. (iii) Experiments on QaTa-COV19-v2 and MosMedData+ show that TeSi achieves competitive performance (91.37% Dice) with interpretable prototype-to-report alignment.

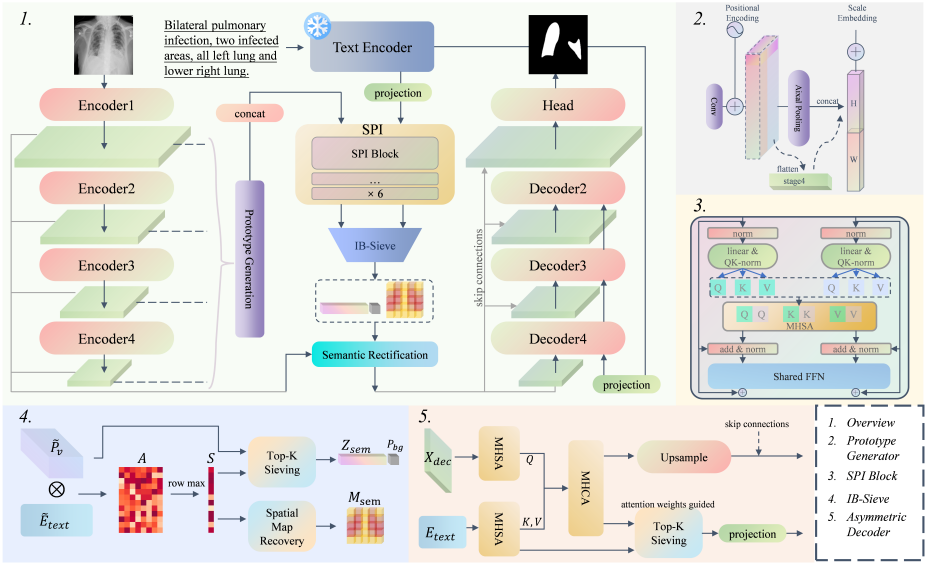

Fig. 1. 1. Overview of TeSi. 2. Axial Prototype Generator compresses visual features. 3. SPI aligns modalities. 4. IB-Sieve ranks and hard-selects visual prototypes using text guidance. 5. Asymmetric Decoder with Adaptive Text Fusion ensures stable reconstruction.

2 Methodology

As illustrated in Fig. 1, TeSi follows a streamlined encoder-decoder architecture. Given an input image, the *Axial Prototype Generator* first compresses multi-scale features into compact prototypes. These prototypes then interact with text embeddings via the *Semantic Prototype Interaction (SPI) module* for alignment. Crucially, the *IB-Sieve Mechanism* ranks and selects informative prototypes early in the encoding phase, suppressing visual redundancy. Finally, the sieved features are reconstructed by an *Asymmetric Decoder*, which employs Adaptive Text Fusion to refine boundaries while maintaining semantic stability.

2.1 Visual Prototype Generation

To address computational inefficiency from visual redundancy, we compress high-resolution feature maps into compact axial prototypes, reducing sequence length from quadratic $H \times W$ to linear $H + W$.

Let the visual backbone (ConvNeXt [18]) output feature maps at four scales $i \in \{1, \ldots, 4\}$. We align the channel dimension to a unified D via a 1×1 convolution and inject 2D absolute position encodings. A learnable scale embedding $E^{(i)}_{scale}$ is added to distinguish scales. For a feature map $X_i \in \mathbb{R}^{H_i \times W_i \times D}$ $(i \neq 4)$, we generate two sets of axial prototypes. For the last scale (7×7), we keep the

dense feature map to preserve the receptive field:

$$P_H^{(i)} = Pool_W(X_i) \in \mathbb{R}^{H_i \times D}, \quad P_W^{(i)} = Pool_H(X_i) \in \mathbb{R}^{W_i \times D}. \tag{1}$$

Concatenating tokens from all scales yields the visual prototype sequence P_v.

2.2 Semantic Prototype Interaction (SPI)

To align modalities in a shared space efficiently, we propose the Semantic Prototype Interaction (SPI) module, avoiding computationally expensive dense cross-attention. We fuse visual prototypes P_v with text embeddings E_{text} (extracted via CXR-BERT [19]).

First, following an initial normalization, we project the visual and textual features using separate matrices ($W_{Q,K,V}^{img}$ and $W_{Q,K,V}^{txt}$), and then apply RMSNorm to the resulting Queries and Keys:

$$Q_{img} = RMS(P_v W_Q^{img}), K_{txt} = RMS(E_{text} W_K^{txt}), V_{txt} = E_{text} W_V^{txt}, \tag{2}$$

$$Q_{text} = RMS(E_{text} W_Q^{text}), K_{img} = RMS(P_v W_K^{img}), V_{img} = P_v W_V^{img}. \tag{3}$$

We concatenate them to form unified sequences $Q_{uni}, K_{uni}, V_{uni}$ and perform Multi-Head Self-Attention (MHSA):

$$Z_{uni} = Softmax\left(\frac{Q_{uni} K_{uni}^T}{\sqrt{D}}\right) V_{uni}. \tag{4}$$

The output Z_{uni} is split back into visual ($\hat{P}_v$) and textual ($\hat{E}_{text}$) components. Both pass through a Weight-Shared Feed-Forward Network (FFN) to enforce semantic alignment:

$$\tilde{P}_v = \hat{P}_v + FFN(LN(\hat{P}_v)), \quad \tilde{E}_{text} = \hat{E}_{text} + FFN(LN(\hat{E}_{text})). \tag{5}$$

2.3 The IB-Sieve Mechanism

The core IB-Sieve mechanism is inspired by the Information Bottleneck principle and operates early in the encoder to filter visual prototypes by textual relevance. It imposes a structural bottleneck on the visual representation (Z) derived from multi-scale features (X), suppressing redundant information while preserving task-relevant semantics for segmentation, rather than explicitly optimizing mutual information objectives.

Relevance Ranking and Hard Selection. We derive the semantic importance score $S \in \mathbb{R}^{|P_v|}$ for each visual prototype by computing the maximal similarity with any text token. Let $\mathcal{A} = \tilde{P}_v(\tilde{E}_{text})^T$ be the affinity matrix, where $\mathcal{A}_{k,j}$ is the similarity between the k-th visual prototype and the j-th text token:

$$S_k = \max_j(\mathcal{A}_{k,j}). \tag{6}$$

We sort prototypes by S and use a fixed selection ratio ($k = 0.5$) to form the high-semantic set $\mathcal{Z}_{sem}$. The remaining low-relevance prototypes form the residual set $P_{rem} = \tilde{P}_v \setminus \mathcal{Z}_{sem}$ and are averaged into a single background token p_{bg} to preserve global context:

$$p_{bg} = \frac{1}{|P_{rem}|} \sum_{p \in P_{rem}} p. \tag{7}$$

This token p_{bg} is appended to $\mathcal{Z}_{sem}$.

Spatial Semantic Map Recovery. To inject textual priors and enable optimization of the non-differentiable sieving process, we reconstruct a 2D attention map from 1D scores via outer product and normalization. For scale i, let $S_H^{(i)}$ and $S_W^{(i)}$ be the importance scores of the height and width prototypes:

$$M_{raw}^{(i)} = S_H^{(i)} \cdot (S_W^{(i)})^T, \quad M_{sem}^{(i)} = \frac{M_{raw}^{(i)} - \min(\cdot)}{\max(\cdot) - \min(\cdot)}. \tag{8}$$

This map is normalized per-scale and concatenated to the rectified feature map $\tilde{X}_i$ as a spatial prior, providing a direct gradient path for relevance ranking.

Semantic Rectification. Finally, refined prototypes are injected back into their corresponding feature levels. For high-resolution scales, a cross-attention layer uses sieved tokens $[\mathcal{Z}_{sem}, p_{bg}]$ as Keys/Values and original X_i as Queries. For the last scale, a 1×1 convolution suffices.

2.4 Asymmetric Decoder with Adaptive Fusion

For reconstruction, we use an asymmetric decoder that treats text as a stable anchor, bypassing SPI to reuse pristine text embeddings and mitigate semantic drift from noisy visual feedback.

To match different decoding stages, we propose Adaptive Text Fusion to condense the token sequence conditioned on the current visual state. Specifically, we perform cross-attention with the decoder visual features as Queries and text embeddings as Keys/Values, i.e., $Q = X_{\text{dec}}$ and $(K, V) = E_{text}$. Given $X_{\text{dec}} \in \mathbb{R}^{H \times W \times C}$ and $E_{text} \in \mathbb{R}^{L \times C}$, we use the resulting attention weights to estimate token relevance, retain [CLS], keep the top-K_t relevant tokens, and fuse the rest into one token via attention-weighted aggregation. The compact sequence (length $K_t + 2$) guides subsequent cross-attention for boundary refinement.

2.5 Objective Function

We use a hybrid loss for training:

$$\mathcal{L}_{total} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{dice}\mathcal{L}_{dice}, \tag{9}$$

where $\mathcal{L}_{ce}$ is Cross-Entropy (pixel-wise classification) and $\mathcal{L}_{dice}$ is Dice loss (class imbalance, shape consistency). Following standard practice [4], we empirically set $\lambda_{ce} = \lambda_{dice} = 1.0$.

3 Experiments

3.1 Datasets

We validate TeSi on two public datasets, used under their open-access licenses. QaTa-COV19-v2 [20] consists of 9,258 X-rays, which we randomly split into 5,716 for training, 1,429 for validation, and 2,113 for testing. MosMedData+ [21] contains 2,729 CT slices, split into 2,183 training, 273 validation, and 273 testing images. We follow the same data split protocol as in prior work (e.g., LViT [9]), and also use the report text annotations provided in those works. We report Dice and mIoU as evaluation metrics.

3.2 Implementation Details

We use PyTorch and MONAI [22] with a single NVIDIA RTX 4090 GPU. We initialize the image encoder with a pre-trained ConvNeXt-tiny [18] and fine-tune it, while keeping the text encoder (pre-trained CXR-BERT [19]) frozen. Images are resized to 224×224. The guidance feature dimension is $D = 384$. In the IB-Sieve, the selection ratio k is set to halve the prototype length. The decoder prunes text length hierarchically ($24 \to 18 \to 12$). We use the AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay 0.01) with Cosine Annealing. The initial learning rate is 6×10^{-5} for QaTa and 8×10^{-5} for MosMed. Batch size is 32. Automatic Mixed Precision (AMP) and early stopping are employed.

3.3 Comparison with Representative Methods

Table 1 compares TeSi with unimodal baselines and recent multimodal methods. Quantitatively, TeSi achieves competitive performance on both datasets. On QaTa-COV19-v2, it reaches 91.37% Dice, consistently outperforming recent multimodal approaches such as ABP [16] and FMISeg [23]. Notably, FMISeg depends on bidirectional frequency-domain convolution branches, which add computational overhead, whereas TeSi attains higher performance by optimizing cross-modal semantic alignment, highlighting the advantage of our strategy over visual-branch architectural expansion. Furthermore, on MosMedData+, despite the limited training data that may constrain attention-based cross-modal learning in SPI, our method maintains competitive performance (77.96% Dice). This underscores the motivation behind our multi-scale prototype design: by capturing both fine-grained lesion details and global anatomical context across different scales, the IB-Sieve mechanism efficiently filters key features even in data-scarce scenarios, avoiding the noise overfitting seen in dense fusion methods (e.g., LViT [9]). Qualitatively, Fig. 2 visually confirms these improvements. For example, in Row 1, prior methods (e.g., LViT, AT) misinterpret report cues, generating false positives in the healthy left lung despite the text explicitly specifying "Unilateral... right lung." In contrast, TeSi yields precise boundaries with significantly fewer false positives, corroborating that our early sieving strategy effectively suppresses task-irrelevant background features at the source before they corrupt the final prediction.

Table 1. Quantitative comparison on QaTa-COV19-v2 and MosMedData+. Baseline results are taken from the original publications, with "–" denoting unavailable entries. Best and second-best results are shown in **bold** and underline.

Method	Modality	QaTa-COV19-v2		MosMedData+	
		Dice (%)	mIoU (%)	Dice (%)	mIoU (%)
U-Net [2]	Image	79.02	69.46	64.60	50.73
U-Net++ [3]	Image	79.62	70.25	71.75	58.39
nnU-Net [4]	Image	80.42	70.81	72.59	60.36
TransUNet [6]	Image	78.63	69.13	71.24	58.44
LViT [9]	Img+Txt	84.92	73.79	74.57	61.33
AT [10]	Img+Txt	89.78	81.45	–	–
TGCAM [11]	Img+Txt	90.60	82.81	77.82	63.69
ABP [16]	Img+Txt	91.03	83.53	–	–
FMISeg [23]	Img+Txt	91.21	83.84	**79.30**	**65.71**
TeSi (Ours)	Img+Txt	**91.37**	**84.12**	77.96	63.87

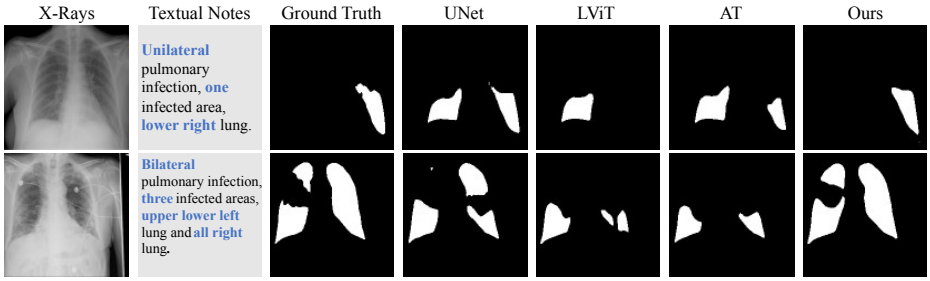

Fig. 2. Visual comparison with representative methods on QaTa-COV19-v2.

3.4 Ablation Study

We validate our design decisions on QaTa-COV19-v2 (Table 2). First, regarding the *Interaction Paradigm*, the Asymmetric Decoder (Row 7) outperforms the Symmetric design (Row 2), validating that controlled interaction prevents semantic drift. Notably, the Encoder-Only variant (Row 3) already yields competitive results, proving the core gain stems from early sieving. Second, regarding *Token Scale*, removing the IB-Sieve mechanism (Row 4) leads to a performance drop, confirming that simply fusing features allows visual redundancy to overwhelm text. Finally, replacing SPI with standard MHCA (Row 5) or removing the Spatial Semantic Map Recovery module (Row 6) results in lower performance, confirming that our specific architectural components are essential for precise alignment and localization.

3.5 Interpretability via Spatial Semantic Maps

To probe *how* the IB-Sieve aligns textual concepts with visual features, we visualize the Spatial Semantic Maps in Fig. 3. These maps reflect simultaneous

Table 2. Ablation analysis. SPI: Semantic Prototype Interaction; Map: Spatial Semantic Map Recovery.

Model Variant	Interaction	Sieving	Decoder	Dice (%)	mIoU (%)
1. Uni-modal Baseline	-	-	CNN	87.93	78.46
2. Symmetric Design	SPI	✓	Symmetric	91.11	83.68
3. Encoder Only	SPI	✓	CNN	91.14	83.73
4. w/o Sieving	SPI	✗(+Map)	Asym.	91.02	83.52
5. w/ Standard Attn	MHCA	✓	Asym.	91.19	83.80
6. w/o Spatial Map	SPI	✓(-Map)	Asym.	91.25	83.91
7. TeSi (Full)	**SPI**	✓	**Asym.**	**91.37**	**84.12**

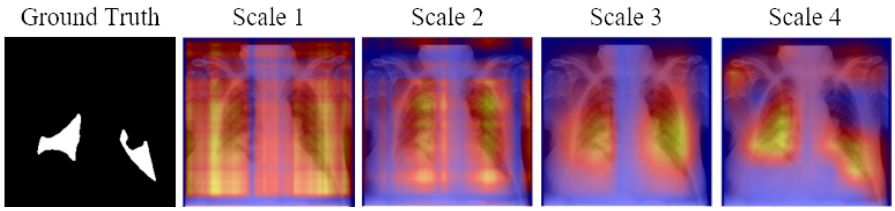

Fig. 3. Recovered Spatial Semantic Maps across four scales, illustrating multi-scale text–visual alignment.

multi-scale interactions in the SPI module rather than a layer-by-layer filtering process. At the high-resolution scale (Scale 1, 2), correlations capture broad, high-frequency edges; at the low-resolution scale (Scale 3, 4), the larger receptive field localizes global lesion boundaries (e.g., "lower right lung"). This contrast shows that our method leverages multi-scale prototypes: shallow scales for fine contours and deep scales for global positioning. Our prototype design also enables text-token-level attention analysis, revealing how specific medical terms trigger spatially distinct activations.

4 Conclusion

In this paper, we identify two key limitations in text-guided medical image segmentation: the *Token Scale* mismatch and a suboptimal *Interaction Paradigm*. Late-stage fusion lets task-irrelevant redundancy propagate and corrupt representations. We propose TeSi, a text-guided prototype sieving framework inspired by the Information Bottleneck principle, which suppresses redundancy early and stabilizes reconstruction. The IB-Sieve ranks and hard-selects text-relevant visual prototypes to block background noise at the source, while the Asymmetric Decoder with Adaptive Text Fusion prevents semantic drift during boundary refinement. Experiments on QaTa-COV19-v2 and MosMedData+ show competitive performance with improved interpretability. Multi-scale visual–text align-

ment yields precise localization and fine-grained contours. A limitation is that our evaluation is restricted to 2D COVID-19 lung datasets, whose homogeneous backgrounds may not fully reflect other organs, MRI, or 3D volumes.

Acknowledgments. This work is supported by National Natural Science Foundation of China (NSFC 62371325), Sichuan Science and Technology Program 2025NSFJQ0050, Sichuan Science and Technology Program 2024ZDZX0018.

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

References

1. Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D.: Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. IEEE Reviews in Biomedical Engineering 14, 4-15 (2020)
2. 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. Springer (2015)
3. 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)
4. Isensee, F., Jaeger, P.F., Kohl, S.A.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)
5. Siddique, N., Sidike, P., Elkin, C.P., Devabhaktuni, V.: U-Net and its variants for medical image segmentation: A review of theory and applications. IEEE Access 9, 82031-82057 (2021)
6. 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)
7. 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)
8. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: International MICCAI Brainlesion Workshop. pp. 272-284. Springer (2021)
9. 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)
10. 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. Springer (2023)
11. 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)

12. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2-25. PMLR (2022)
13. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.-J.: DynamicViT: Efficient vision transformers with dynamic token sparsification. Advances in Neural Information Processing Systems 34, 13937-13949 (2021)
14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: European Conference on Computer Vision. pp. 38-55. Springer (2024)
15. Jeon, Y.S., Yang, H., Fu, H., Kway, Y., Feng, M.: No more sliding window: efficient 3D medical image segmentation with differentiable top-k patch sampling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 376-386. Springer (2025)
16. Zeng, X., Zeng, P., Cui, J., Li, A., Liu, B., Wang, C., Wang, Y.: ABP: Asymmetric bilateral prompting for text-guided medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 54-64. Springer (2024)
17. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)
18. 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. IEEE (2022)
19. 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. Springer (2022)
20. Degerli, A., Ahishali, M., Kiranyaz, S., Chowdhury, M.E.H., Gabbouj, M.: Reliable COVID-19 detection using chest X-ray images. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 185-189. IEEE (2021)
21. Morozov, S.P., Andreychenko, A.E., Pavlov, N.A., Vladzymyrskyy, A.V., Ledikhova, N.V., Gombolevskiy, V.A., Blokhin, I.A., Gelezhe, P.B., Gonchar, A.V., Chernina, V.Y.: MosMedData: Chest CT scans with COVID-19 related findings dataset. arXiv preprint arXiv:2005.06465 (2020)
22. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: MONAI: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
23. Yu, B., Yang, J., Du, Z., Huang, Y., Li, C., Wang, L.: Frequency-domain multimodal fusion for language-guided medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 278-288. Springer (2025)