Text-Guided Multi-Frequency Latent Diffusion for Medical Image Segmentation

Qiang Gao[1][0000−0002−1102−8784], Yi Wang[2]*, Yong Zhang[2], Lan Du[1], Yong Li[2], and Cunjian Chen[1]*

[1] Department of Data Science and AI, Monash University, Melbourne, Australia
[2] College of Computer Science, Chongqing University, Chongqing, China
* corresponding author: yiwang@cqu.edu.cn, Cunjian.Chen@monash.edu

Abstract. In latent diffusion-based segmentation, the model is trained to recover anatomically consistent regions from noisy latent mask representations conditioned on the input image. Recent diffusion-based approaches provide strong generative priors, but their conditioning is often limited to image or latent features, without explicit structural or semantic guidance. This may lead to unstable predictions and imprecise boundary delineation. In this work, we propose a text-guided multi-frequency latent diffusion framework for medical image segmentation, termed TMF-Seg. The model introduces an Image–Text Conditional Fusion module that encodes image and text conditions and integrates them through cross-attention to obtain a semantically informed representation. We further design a Text-Guided Multi-Frequency Enhancement module that decomposes the fused features into high-, mid-, and low-frequency components and adaptively reweights them according to the textual context, providing complementary semantic and structural priors for the diffusion process. The enhanced conditional features are then fed into a latent diffusion U-Net for efficient single-step segmentation. Experiments on three public datasets, Kvasir-SEG, MosMedData+, and QaTa-COV19, show that TMF-Seg consistently improves segmentation accuracy over several strong baselines. The code is publicly available at https://github.com/qczggaoqiang/TMF-Seg.

Keywords: Medical Image Segmentation · Latent Diffusion · Image–Text Conditional Fusion · Text-Guided Multi-Frequency Enhancement.

1 Introduction

Medical image segmentation plays a fundamental role in many clinical applications, including disease diagnosis [3], treatment planning [21], and surgical guidance [2]. Accurate delineation of anatomical structures or lesions is essential for reliable quantitative analysis and downstream decision-making [6]. With the rapid development of deep learning, convolutional and transformer-based models have become the mainstream solutions, achieving strong performance by learning hierarchical and global representations [17,11]. However, target regions in medical images often exhibit low contrast, ambiguous boundaries, and complex

anatomical structures [9,20]. Under such conditions, relying solely on image-driven features is insufficient, and the absence of explicit structural or semantic priors can lead to unstable and anatomically inconsistent segmentation results.

Recently, diffusion-based approaches have been introduced into medical image segmentation [22,23,16,5,13]. In latent diffusion-based segmentation, the model is trained to recover anatomically consistent regions from noisy latent mask representations conditioned on the input image, benefiting from strong generative priors and improved robustness [12,8]. Despite these advantages, the performance of diffusion segmentation is largely determined by the conditioning mechanism [24]. Most existing methods rely solely on image or latent features as conditions [7,14], without explicitly modeling structural or semantic guidance. As a result, the denoising process may lack clear signals about boundary sharpness, global structure, or task-level semantics, which can lead to unstable predictions and suboptimal delineation in challenging regions.

To address these limitations, we propose a **t**ext-guided **m**ulti-**f**requency latent diffusion framework for medical image **seg**mentation, termed TMF-Seg, which enhances the diffusion conditioning with both semantic and structural priors. Instead of modifying the diffusion backbone, our method focuses on redesigning the conditional encoder to provide both semantic and structural priors. While diffusion-based segmentation has shown promising results, explicit integration of textual semantics and frequency-aware structural modeling remains underexplored. Specifically, we introduce an Image–Text Conditional Fusion (ITCF) module that integrates visual and textual information through cross-attention, enabling the model to incorporate task-level semantic cues. We further propose a Text-Guided Multi-Frequency Enhancement (TGME) module, which decomposes the fused features into multiple frequency components and adaptively reweights them according to the textual context. This design performs text-guided frequency modulation, enabling adaptive emphasis on fine details, boundary structures, and global anatomical context, and thus provides complementary semantic and structural guidance for the diffusion process. We evaluate the proposed method on three public medical image segmentation datasets, including Kvasir-SEG, MosMedData+, and QaTa-COV19 [25]. Experimental results demonstrate that TMF-Seg consistently improves segmentation accuracy over several strong baselines.

Motivated by the limitations discussed above, the main contributions of this work can be summarized as follows:

1. We propose a text-guided multi-frequency latent diffusion framework, termed **TMF-Seg**, which enhances diffusion conditioning with both semantic and structural priors for medical image segmentation.
2. We introduce an **Image–Text Conditional Fusion** module that integrates visual and textual information to produce semantically informed conditional representations.
3. We design a **Text-Guided Multi-Frequency Enhancement** module that adaptively reweights frequency components based on textual context, im-

proving structural consistency and segmentation accuracy across multiple datasets.

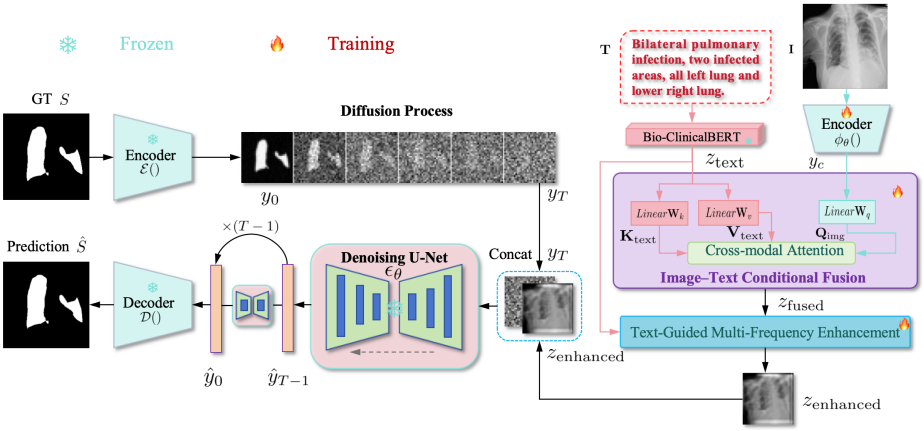

Fig. 1. Overview of the proposed **TMF-Seg** framework. Image and text features are fused via the Image–Text Conditional Fusion (ITCF) module and enhanced by the Text-Guided Multi-Frequency Enhancement (TGME) module. The resulting condition guides a latent diffusion U-Net to generate the final segmentation.

2 Methodology

Figure 1 illustrates the overall architecture of the proposed TMF-Seg. Given a medical image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and its textual description $\mathbf{T}$, the framework generates a segmentation mask $\hat{S}$ through a conditional latent diffusion process. The image is first encoded into a latent representation $y_c = \phi_\theta(I)$ using a pre-trained VAE encoder [12], while the text description is processed by a frozen Bio-ClinicalBERT [1] to obtain semantic features z_{text}. These two modalities are fused via a Image–Text Conditional Fusion (ITCF) module to produce semantically informed features z_{fused}, which are further refined by a Text-Guided Multi-Frequency Enhancement (TGME) module to obtain the enhanced condition z_{enhanced}. The diffusion process is then carried out in the latent space of a pre-trained segmentation autoencoder: the Ground-Truth (GT) mask S is first encoded as a latent representation $y_0 = \mathcal{E}(S)$, which is then progressively perturbed with Gaussian noise to obtain a noisy latent y_T at timestep T. The denoising network ϵ_θ is trained to recover the clean representation from y_T conditioned on z_{enhanced}, producing the predicted latent $\hat{y}_0$, which is then decoded by the segmentation decoder $\mathcal{D}$ to obtain the final prediction $\hat{S}$. The autoencoder $(\mathcal{E}, \mathcal{D})$ and the diffusion model remain frozen, while the proposed condition encoder is trained.

2.1 Image–Text Conditional Fusion

To incorporate semantic guidance into the denoising process, we design an Image-Text Conditional Fusion module that integrates visual and textual representations in the latent space. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ and its textual description $\mathbf{T}$, we encode them using separate encoders. The image is mapped by a pre-trained VAE encoder ϕ_θ to a spatial latent feature $y_c \in \mathbb{R}^{C\times h\times w}$, where $C = 4$ and $h = w = 32$ for 256×256 inputs. The text is embedded by a frozen Bio-ClinicalBERT [1], producing contextual representations $z_{\text{text}} \in \mathbb{R}^{L\times D}$, where $L = 10$ is the sequence length and $D = 768$ is the embedding dimension. We employ a cross-attention mechanism [6] to fuse the two modalities:

$$z_{\text{fused}} = \text{Cross-modal Attention}(\mathbf{Q}_{\text{img}}, \mathbf{K}_{\text{text}}, \mathbf{V}_{\text{text}}), \tag{1}$$

where $\mathbf{Q}_{\text{img}} = \mathbf{W}_q y_c$ denotes queries derived from image features, and $\mathbf{K}_{\text{text}} = \mathbf{W}_k z_{\text{text}}$, $\mathbf{V}_{\text{text}} = \mathbf{W}_v z_{\text{text}}$ are keys and values projected from text embeddings. This design enables spatial image features to attend to semantically relevant textual cues, yielding a semantically informed conditional representation that serves as guidance for the denoising U-Net.

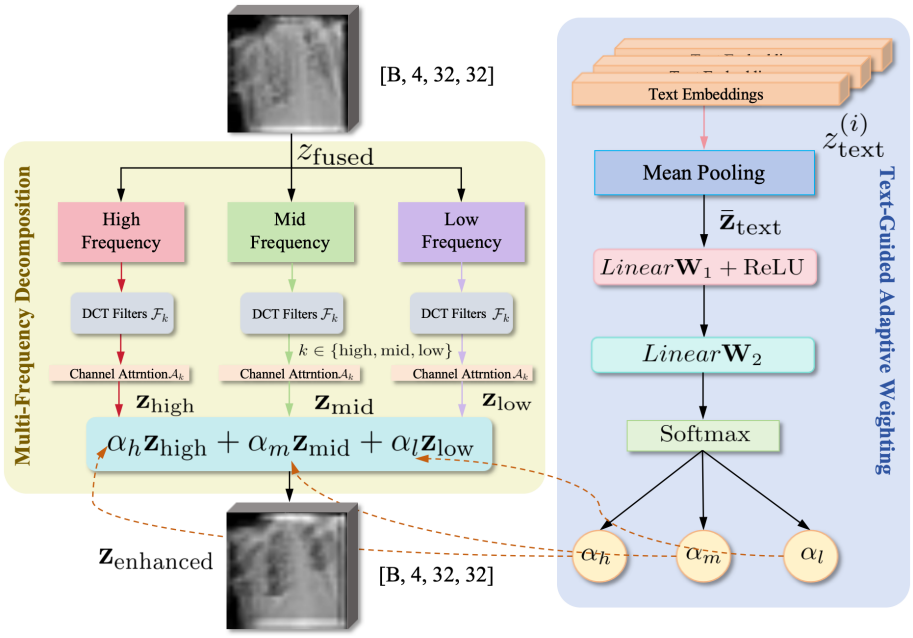

Fig. 2. Overview of the proposed Text-Guided Multi-Frequency Enhancement module.

2.2 Text-Guided Multi-Frequency Enhancement

After cross-modal fusion, we introduce a Text-Guided Multi-Frequency Enhancement module, as illustrated in Fig. 2, to refine the fused latent features by incorporating structural priors in the frequency domain. Medical image segmentation requires capturing information at different spatial frequencies, including fine-grained details, structural boundaries, and global anatomical context. Given the fused feature $z_{\text{fused}} \in \mathbb{R}^{B\times C\times H\times W}$, we perform DCT-based frequency decomposition [15] and construct three parallel branches corresponding to high-, mid-, and low-frequency components. Each branch applies multi-frequency channel attention (MFCA) [15] to enhance the fused features by selectively emphasizing frequency-specific components:

$$\mathbf{z}_k = z_{\text{fused}} \odot \sigma(\mathcal{A}_k(\mathcal{F}_k(z_{\text{fused}}))), \tag{2}$$

where $k \in \{\text{high}, \text{mid}, \text{low}\}$, $\mathcal{F}_k$ denotes DCT filtering with pre-defined frequency indices, $\mathcal{A}_k$ is a lightweight channel attention module, σ is the sigmoid function, and $\odot$ represents element-wise multiplication.

Text-Guided Adaptive Weighting. To adaptively modulate the contribution of each frequency band, we compute a global text representation:

$$\bar{\mathbf{z}}_{\text{text}} = \frac{1}{L}\sum_{i=1}^{L} z_{\text{text}}^{(i)}, \tag{3}$$

where L is the text sequence length and $z_{\text{text}}^{(i)}$ denotes the i-th token embedding. This global representation is then mapped to normalized frequency weights through a two-layer MLP ($\mathbf{W}_1$, $\mathbf{W}_2$) with ReLU activation and softmax normalization:

$$[\alpha_h, \alpha_m, \alpha_l] = \text{Softmax}(\mathbf{W}_2\,\text{ReLU}(\mathbf{W}_1\bar{\mathbf{z}}_{\text{text}})), \tag{4}$$

where $\alpha_h, \alpha_m, \alpha_l \in [0, 1]$ are the predicted importance weights for high-, mid-, and low-frequency components, respectively, with $\alpha_h + \alpha_m + \alpha_l = 1$.

The enhanced representation is obtained by:

$$\mathbf{z}_{\text{enhanced}} = \alpha_h\mathbf{z}_{\text{high}} + \alpha_m\mathbf{z}_{\text{mid}} + \alpha_l\mathbf{z}_{\text{low}}. \tag{5}$$

This design enables semantic-aware frequency modulation, allowing the diffusion model to receive both structural and semantic guidance. During training, we follow SDSeg [12] and employ the same latent diffusion training objective without modification.

3 Experiments

3.1 Experiments details

Our model is trained on a single NVIDIA RTX 3090 GPU with 24GB memory. The model is optimized for 100,000 steps using the AdamW optimizer with a

Table 1. Quantitative comparison on Kvasir-SEG, MosMedData+, and QaTa-COV19. The best results are shown in **bold**.

Method	Backbone	Kvasir-SEG		MosMedData+		QaTa-COV19	
		Dice ↑	mIoU ↑	Dice ↑	mIoU ↑	Dice ↑	mIoU ↑
U-Net [18]	CNN	78.94	70.21	64.60	50.73	79.02	69.46
U-Net++ [26]	CNN	80.35	71.85	71.75	58.39	79.62	70.25
TransUNet [4]	CNN–Transformer	82.05	74.12	71.24	58.44	78.63	69.13
LViT [10]	CNN–Transformer	86.54	78.33	74.57	61.33	83.66	75.11
TSLDeg [24]	Latent Diffusion	88.72	81.45	77.83	63.92	85.41	77.36
LEAF [8]	Latent Diffusion	89.65	82.97	78.94	64.88	86.73	79.92
SDSeg [12]	Latent Diffusion	91.03	85.80	80.97	65.68	89.92	81.85
TMF-Seg (Ours)	Latent Diffusion	**92.63**	**86.06**	**81.34**	**67.21**	**91.42**	**82.43**

base learning rate of 1×10^{-5} and a batch size of 8. We adopt a KL-regularized autoencoder and latent diffusion model with a downsampling rate of $r = \frac{H}{h} = \frac{W}{w} = 8$. The model takes RGB images of size 256×256 as input, producing latent representations of size 32×32 with 4 channels. The diffusion backbone and autoencoder are initialized from pre-trained Stable Diffusion weights, while the additional parameters in the denoising U-Net for concatenated inputs are initialized to zero.

3.2 Datasets and Evaluation Metrics

To evaluate the proposed method TMF-Seg, we conduct experiments on three public medical image segmentation datasets, including Kvasir-SEG, MosMedData+, and QaTa-COV19 [25]. For MosMedData+ and QaTa-COV19, we use the released text annotations from LViT [10]; for Kvasir-SEG, the corresponding text annotations are available upon reasonable request by email. The Kvasir-SEG dataset contains 1,000 colonoscopy images with corresponding polyp segmentation masks. Following common practice, we randomly split the dataset into 800 training, 100 validation, and 100 test images. The MosMedData+ dataset includes 2,729 COVID-19 lung CT slices annotated with binary infection masks, which are partitioned into 2,183 training, 273 validation, and 273 test samples. The QaTa-COV19 dataset contains 9,258 chest X-ray images with COVID-19 lesion annotations, and is split into 5,716 training, 1,429 validation, and 2,113 test samples. Following common practice in medical image segmentation, we adopt two widely used evaluation metrics, namely the Dice coefficient (Dice) and mean intersection-over-union (mIoU), to quantitatively assess the segmentation performance.

3.3 Comparison With State-of-the-Art Methods

We compare TMF-Seg with several representative segmentation approaches, including classical convolutional models (U-Net [18], U-Net++ [26]), hybrid

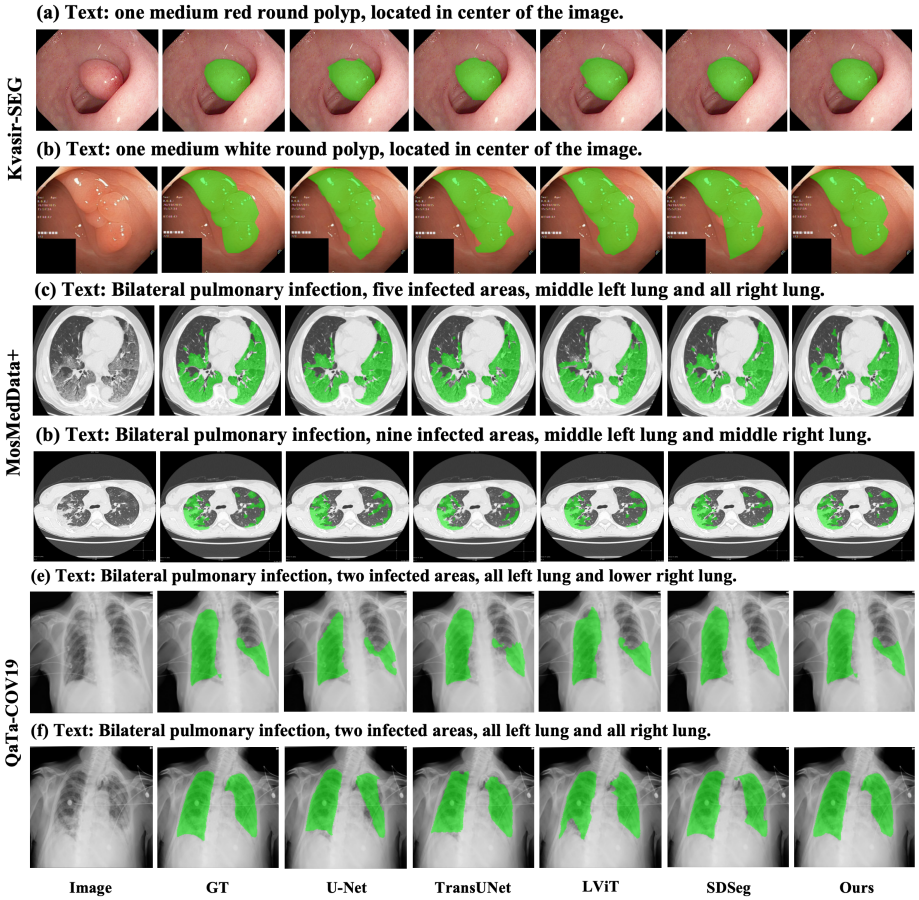

Fig. 3. Qualitative comparison on three datasets. Top: Kvasir-SEG; Middle: MosMedData+; Bottom: QaTa-COV19. From left to right in each row: input image, ground truth (GT), U-Net, TransUNet, LViT, SDSeg, and TMF-Seg (Ours).

CNN–Transformer architectures (TransUNet [4], LViT [10]), recent latent diffusion-based segmentation methods (TSLDeg [24], LEAF [8], SDSeg [12]). These methods cover different architectural paradigms, enabling a comprehensive evaluation of our approach. Quantitative results on Kvasir-SEG, MosMedData+, and QaTa-COV19 are reported in Table 1. TMF-Seg consistently achieves superior Dice and mIoU scores compared to the competing methods across all three datasets. In particular, compared with the latent diffusion baseline SDSeg, our method demonstrates consistent improvements, highlighting the effectiveness of the proposed ITCF and TGME modules. Qualitative comparisons are shown in Fig. 3. TMF-Seg produces more coherent and anatomically consistent segmentation maps, especially in challenging regions with ambiguous boundaries or subtle lesion structures.

Table 2. Ablation study of the proposed modules on QaTa-COV19.

Model	ITCF	TGME	Dice ↑	mIoU ↑
SDSeg (baseline)			89.92	81.85
+ ITCF	✓		90.78	82.18
+ TGME		✓	90.56	82.06
TMF-Seg (full)	✓	✓	**91.42**	**82.43**

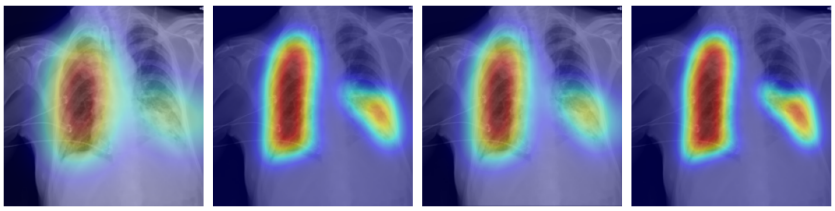

Fig. 4. Grad-CAM style visualizations under different ablation settings on QaTa-COV19. From left to right: SDSeg baseline, +ITCF, +TGME, and the full TMF-Seg model. The introduction of ITCF and TGME progressively refines the model's focus on lesion-relevant and structure-aware regions.

3.4 Ablation Study

To evaluate the effectiveness of each proposed component, we conduct ablation experiments on QaTa-COV19 by progressively introducing the ITCF and TGME modules into the SDSeg baseline. Table 2 reports the quantitative results. Starting from the original SDSeg framework, incorporating ITCF leads to consistent improvements, demonstrating the benefit of integrating textual semantic information through cross-modal fusion. Further adding TGME yields additional gains, indicating that text-guided frequency modulation provides complementary structural priors for the diffusion process. The full TMF-Seg model achieves the highest performance on QaTa-COV19 among the ablation variants, confirming the effectiveness and complementarity of the proposed components.

To provide further insight, we adopt gradient-based localization (Grad-CAM) visualizations to illustrate how the model's salient regions change under different ablation settings (SDSeg, +ITCF, +TGME, and the full TMF-Seg) [19], as shown in Fig. 4. The visualizations indicate that introducing ITCF and TGME progressively guides the model to attend to more lesion-relevant and structure-aware regions, while the full model exhibits the most concentrated and coherent responses. These qualitative observations are consistent with the quantitative improvements reported in Table 2.

4 Conclusion

In this work, we presented TMF-Seg, a text-guided multi-frequency latent diffusion framework for medical image segmentation. Built upon a latent diffusion

backbone, the proposed method enhances the conditioning mechanism through Image–Text Conditional Fusion and Text-Guided Multi-Frequency Enhancement, enabling structure-aware frequency modulation in the latent space. By integrating clinical textual information and adaptively reweighting frequency components, TMF-Seg provides complementary semantic and structural guidance for the denoising process. Experiments on Kvasir-SEG, MosMedData+, and QaTa-COV19 demonstrate consistent improvements over strong baselines, validating the effectiveness of the proposed design. These findings suggest that enriching diffusion conditioning with textual semantics and structure-aware frequency priors can improve robustness and anatomical consistency in medical image segmentation.

Acknowledgements. This work was supported by a key joint project of Chongqing Health Commission and Science and Technology Bureau (2024ZDXM007), Central University basic research young teachers and students research ability promotion subproject (2023CDJYGRH-ZD06), and Chongqing Technology Innovation and Application Development Project (CSTB2023TIAD-KPX0050).

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

References

1. 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)
2. Avrunin, O., Alkhorayef, M., Saied, H.F.I., Tymkovych, M.: The surgical navigation system with optical position determination technology and sources of errors. Journal of Medical Imaging and Health Informatics **5**(4), 689–696 (2015)
3. Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., Jose, A., Roy, R., Merhof, D.: Advances in medical image analysis with vision transformers: a comprehensive review. Medical Image Analysis **91**, 103000 (2024)
4. 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)
5. Chowdary, G.J., Yin, Z.: Diffusion transformer u-net for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 622–631. Springer (2023)
6. Feng, C.M.: Enhancing label-efficient medical image segmentation with text-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 253–262. Springer (2024)
7. Hejrati, B., Banerjee, S., Glide-Hurst, C., Dong, M.: Conditional diffusion model with spatial attention and latent embedding for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 202–212. Springer (2024)
8. Huang, Q., Lin, T., Chen, Z., Zheng, F.: Leaf: Latent diffusion with efficient encoder distillation for aligned features in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 384–393. Springer (2025)

9. Lei, M., Wu, H., Lv, X., Wang, X.: Condseg: A general medical image segmentation framework via contrast-driven feature enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4571–4579 (2025)
10. 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)
11. Liang, Z., Zhao, K., Liang, G., Li, S., Wu, Y., Zhou, Y.: Maxformer: Enhanced transformer for medical image segmentation with multi-attention and multi-scale features fusion. Knowledge-Based Systems **280**, 110987 (2023)
12. Lin, T., Chen, Z., Yan, Z., Yu, W., Zheng, F.: Stable diffusion segmentation for biomedical images with single-step reverse process. In: International conference on medical image computing and computer-assisted intervention. pp. 656–666. Springer (2024)
13. Liu, T., Zhang, M., Liu, L., Zhong, J., Wang, S., Piao, Y., Lu, H.: Cridiff: Crisscross injection diffusion framework via generative pre-train for prostate segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 102–112. Springer (2024)
14. Liu, Y., Feng, Y., Cheng, J., Zhan, H., Zhu, Z.: Mambadiff: Mamba-enhanced diffusion model for 3d medical image segmentation. IEEE Transactions on Image Processing (2025)
15. Nam, J.H., Syazwany, N.S., Kim, S.J., Lee, S.C.: Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11480–11491 (2024)
16. Rahman, A., Valanarasu, J.M.J., Hacihaliloglu, I., Patel, V.M.: Ambiguous medical image segmentation using diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11536–11546 (2023)
17. Rahman, M.M., Munir, M., Marculescu, R.: Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11769–11779 (2024)
18. 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)
19. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. International journal of computer vision **128**(2), 336–359 (2020)
20. Shan, D., Li, Z., Li, Y., Li, Q., Tian, J., Hong, Q.: Stpnet: Scale-aware text prompt network for medical image segmentation. IEEE Transactions on Image Processing (2025)
21. Sherer, M.V., Lin, D., Elguindi, S., Duke, S., Tan, L.T., Cacicedo, J., Dahele, M., Gillespie, E.F.: Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiotherapy and Oncology **160**, 185–191 (2021)
22. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 6030–6038 (2024)
23. Yan, P., Li, M., Zhang, J., Li, G., Jiang, Y., Luo, H.: Cold segdiffusion: A novel diffusion model for medical image segmentation. Knowledge-Based Systems **301**, 112350 (2024)

24. Yang, Z., Li, C., Ma, J.: Tsldseg: A texture-aware and semantic-enhanced latent diffusion model for medical image segmentation. Pattern Recognition p. 112795 (2025)
25. Ye, S., Naseem, U., Meng, M., Kim, J.: Alleviating textual reliance in medical language-guided segmentation via prototype-driven semantic approximation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22316–22326 (2025)
26. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: U-net++: A nested u-net architecture for medical image segmentation. In: International workshop on deep learning in medical image analysis. pp. 3–11. Springer (2018)