Hierarchical Text-Guided Brain Tumor Segmentation via Sub-Region-Aware Prompts

Bahram Mohammadi[1,✉], Ta Duc Huy[2], Afrouz Sheikholeslami[1], Qi Chen[2], Vu Minh Hieu Phan[2], Sam White[2], Minh-Son To[3], Xuyun Zhang[1], Amin Beheshti[1], Luping Zhou[4], and Yuankai Qi[1,⋆,✉]

[1] Macquarie University, Sydney, NSW, Australia
[2] Adelaide University, Adelaide, SA, Australia
[3] Flinders University, Adelaide, SA, Australia
[4] University of Sydney, Sydney, NSW, Australia
mohammadibahram71@gmail.com, yuankai.qi@mq.edu.au

Abstract. Brain tumor segmentation remains challenging because the three standard sub-regions, *i.e.*, whole tumor (WT), tumor core (TC), and enhancing tumor (ET), often exhibit ambiguous visual boundaries. Integrating radiological description texts with imaging is promising; however, most multimodal approaches typically compress a report into a single global text embedding shared across all sub-regions, overlooking their distinct clinical characteristics. We propose TextSCP (text-modulated soft cascade architecture with prompts), a hierarchical text-guided framework built on the TextBraTS baseline with three novel components: (1) a text-modulated soft cascade decoder that predicts WT→TC →ET in a coarse-to-fine manner consistent with their anatomical hierarchy. (2) sub-region-aware prompt tuning, using learnable soft prompts with a LoRA-adapted BioBERT encoder to generate specialized representations for each sub-region; (3) text-semantic channel modulators that convert these representations into channel-wise refinement signals, enabling the decoder to emphasize features aligned with clinically described patterns. Results on the TextBraTS dataset demonstrate consistent improvements across all sub-regions, outperforming state-of-the-art methods by 1.7% and 6.2% on Dice and HD95, respectively. Code is available at: https://github.com/Bahram-Mohammadi/TextSCP.

Keywords: Tumor Segmentation · Soft Cascade · Prompt Tuning

1 Introduction

Deep learning has made substantial progress on brain tumor segmentation, with architectures such as U-Net [16], nnU-Net [9], and Swin-UNETR [4] achieving strong performance on the BraTS benchmarks [17]. However, these methods operate exclusively on imaging data, ignoring the rich clinical context available in radiological reports. In practice, radiologists produce free-text annotations that

⋆ Corresponding author

describe the tumor's location, size, enhancement pattern, surrounding edema, and relationship to eloquent structures, which informs the distinction between sub-regions. For instance, the phrase "*ring-enhancing mass with central necrosis*" immediately constrains the spatial relationship between ET and the non-enhancing core, a distinction that is often ambiguous from imaging alone.

The recently introduced TextBraTS [17] benchmark provides the first publicly available dataset that pairs volumetric brain MRI with radiological text descriptions, and a baseline model that fuses SwinTransformer features and BioBERT [10] embeddings via cross-attention. This benchmark shows the value of incorporating clinical language into tumor segmentation. Building on this idea, several text-guided segmentation frameworks have emerged [15,11,17,12,19,18]. VoxTell [15] maps free-form prompts to volumetric masks, achieving strong zero-shot performance. In [19], texts are integrated into a SAM-inspired decoder for 3D medical segmentation. TVPNet [18] leverages text–vision prompt interactions to highlight small or clinically significant structures. Despite this progress, current methods show two key limitations for multi-region tumor segmentation. First, most models use a single shared output head, ignoring the anatomical hierarchy (ET $\subset$ TC $\subset$ WT) and often producing inconsistent predictions, such as ET appearing outside TC. Second, they typically compress the entire report into a single global text embedding, overlooking the dependency of sub-regions on different clinical cues (*e.g.*, edema vs. necrosis vs. enhancement).

To address these limitations, we propose **TextSCP**, a unified hierarchical text-guided framework for brain tumor segmentation. First, to exploit the anatomical hierarchy, we design a text-modulated soft cascade with three prediction branches: WT generates a soft spatial attention prior that guides TC, and TC in turn constrains ET, explicitly encoding the structure ET $\subset$ TC $\subset$ WT. Second, to overcome the limitation of the single global text embedding, TextSCP introduces sub-region-aware prompt tuning, where distinct soft prompts steer a LoRA-adapted BioBERT encoder to produce specialized text representations for WT, TC, and ET. Third, each branch further incorporates a text-semantic channel modulator, which injects its branch-specific text representation into decoder features through a squeeze-and-excitation mechanism, enabling channel-wise refinement aligned with clinically described patterns. These components together form a parameter-efficient yet expressive multi-modal architecture. Our main contributions are summarized as follows:

- We introduce a text-modulated soft cascade decoder with three independent prediction branches that encode the anatomical hierarchy (ET $\subset$ TC $\subset$ WT), where coarser predictions guide finer sub-regions and text modulators inject essential information at each cascade level.
- We propose sub-region-aware prompt tuning, in which distinct learnable soft prompts are prepended to a shared LoRA-adapted BioBERT encoder to produce three branch-specialized text representations.
- The results on the TextBraTS dataset demonstrate consistent enhancement over the baseline. Specifically, average Dice and HD95 are improved by 1.7% and 6.2%, respectively.

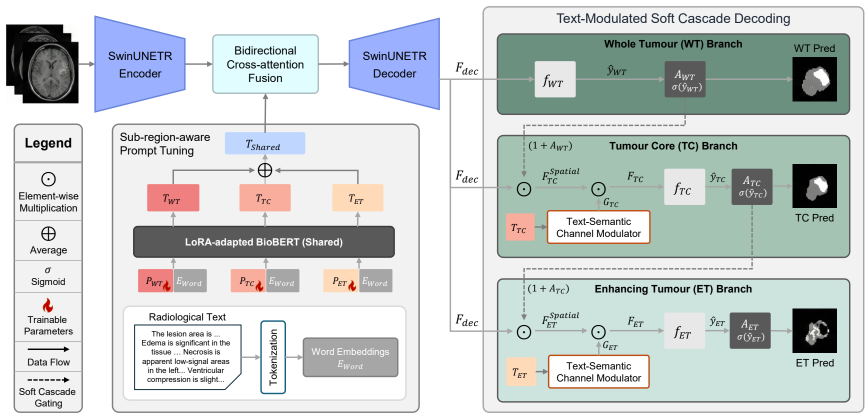

Fig. 1. Main architecture of our TextSCP, which contains three key components: (1) a text-modulated soft cascade decoder that predicts WT→TC→ET in a coarse-to-fine manner (Sec. 2.2); (2) sub-region-aware prompt tuning (Sec. 2.3), which uses learnable soft prompts with a LoRA-adapted BioBERT encoder to generate specialized text representations tailored to each sub-region; and (3) text-semantic channel modulators that convert these representations into channel-wise refinement signals, enabling the decoder to emphasize features aligned with clinically described patterns (Sec. 2.4).

2 Method

We propose TextSCP, a unified hierarchical text-guided framework for brain tumor segmentation (Fig. 1). Built upon the TextBraTS baseline, TextSCP introduces three novel components. First, we design a *soft cascade decoding strategy* with three separate sub-region prediction heads (Sec. 2.2) that explicitly enforce the anatomical hierarchy (ET $\subset$ TC $\subset$ WT). Second, we design a *Sub-Region-Aware Text Adaptation*, which combines parameter-efficient fine-tuning (LoRA [6]) with sub-region-aware prompt tuning to produce branch-specialized text representations, overcoming the limitation of a single global embedding (Sec. 2.3). Third, we design *text-derived semantic channel modulators* that inject branch-specific linguistic priors into each cascade branch (Sec. 2.4).

2.1 Baseline Architecture

The baseline TextBraTS framework consists of a 3D Swin Transformer encoder that extracts hierarchical visual features at five resolutions from four-channel MRI volumes (T1, T1ce, T2, FLAIR), a text encoder from BioBERT that produces embeddings of radiological description text, a sequential cross-attention fusion module at the encoder bottleneck that enables bidirectional reasoning between image and text, and a U-Net decoder with skip connections that reconstructs high-resolution feature maps. These features are passed to a single output head predicting all three tumor sub-regions. We design three new components, detailed below, that largely improve this baseline.

2.2 Soft Cascade Decoding with Separate Sub-Region Heads

The baseline model may fail to exploit the anatomical hierarchy among brain tumor sub-regions because of using a single shared output head. We propose a soft cascade of three independent prediction heads, where coarser regions generate a spatial attention map for finer sub-regions. Let $F_{dec} \in \mathbb{R}^{B\times C\times D\times H\times W}$ denote the high-resolution decoder output features, where $C = 48$. The cascade proceeds through three sequential branches, each equipped with its own convolutional output head:

WT Branch. The WT branch operates directly on the decoder features to predict the binary distinction between tumor and healthy tissue: $\hat{y}_{WT} = f_{WT}(F_{dec})$, $A_{WT} = \sigma(\hat{y}_{WT})$, where f_{WT} is the WT output head and $\sigma(.)$ is the sigmoid function. The resulting soft attention map $A_{WT} \in \mathbb{R}^{B\times 1\times D\times H\times W}$ provides a voxel-wise probabilistic prior over the tumor location, which is propagated to the subsequent TC branch.

TC and ET Branch. The TC branch incorporates spatial guidance from the WT prediction. The decoder features are modulated by the WT attention gate to focus computation on the predicted tumor region: $F^{spatial}_{TC} = F_{dec} \odot (1+A_{WT})$, where $\odot$ denotes element-wise multiplication. The term $(1+A_{WT})$ implements a soft residual gating mechanism: voxels within the predicted tumor region receive amplified features (up to $2\times$ when $A_{WT} \to 1$), while voxels outside the tumor retain their original features ($1\times$ when $A_{WT} \to 0$). This avoids information loss inherent in hard masking while still providing a strong spatial inductive bias.

The spatially gated features are then subject to text-based semantic modulation (described in Sec. 2.4), resulting in the final TC features, from which the TC logit is predicted: $F_{TC} = F^{spatial}_{TC} \odot G_{TC}$, $\hat{y}_{TC} = f_{TC}(F_{TC})$. The TC attention gate is then computed as $A_{TC} = \sigma(\hat{y}_{TC})$ and propagated to the next stage, *i.e.*, ET branch. The ET logits $\hat{y}_{ET}$ are predicted in the same manner as in the previous stage.

2.3 Sub-Region-Aware Text Adaptation

To overcome the limitation of a single shared text embedding, we propose *Sub-Region-Aware Text Adaptation*. It comprises two components detailed below.

Parameter-Efficient Text Encoding via LoRA We propose using LoRA to fine-tune the text encoder, which learns task-specific representations that better capture the discriminative semantics required for sub-region differentiation. We apply LoRA specifically to the query and value projection matrices. The key projections and all feed-forward layers remain frozen, resulting in only a small number of trainable parameters. For a pretrained weight matrix $W \in \mathbb{R}^{d\times d}$, the adapted forward pass is $Wx + \frac{\alpha}{r}$, where r is the rank and α is the scaling factor.

Sub-Region-Aware Prompt Tuning Each tumor sub-region relies on various linguistic cues: edema descriptors for WT, necrosis-related terms for TC, and enhancement patterns for ET. Hence, we introduce sub-region-aware prompt tuning, a lightweight mechanism that utilizes a single shared LoRA-adapted BioBERT model to produce three branch-specialized representations. We define three sets of learnable soft prompt embeddings $\mathbf{P}_{\mathrm{WT}}$, $\mathbf{P}_{\mathrm{TC}}$, $\mathbf{P}_{\mathrm{ET}} \in \mathbb{R}^{K\times d}$, with prompt length $K = 4$ and hidden dimension $d = 768$. For each sub-region $s \in \{\mathrm{WT}, \mathrm{TC}, \mathrm{ET}\}$, the prompts are prepended to word embeddings and processed via the LoRA-adapted encoder: $\mathbf{T}_s = \mathrm{BioBERT}_{\mathrm{LoRA}}\big(\mathbf{E}_s, \mathbf{M}_s\big)\big[:,\ K:\big]$, where $\mathbf{M}_s$ is the extended attention mask, $\mathbf{E}_s = [\,\mathbf{P}_s; \mathbf{E}_{\mathrm{word}}\,]$, $\mathbf{E}_{\mathrm{word}} \in \mathbb{R}^{L\times d}$ contains the tokenized word embeddings, and the slicing discards prompt-position outputs. This text encoder is shared across the three forward passes, and only the prompt embeddings differ, acting as lightweight steering vectors that condition the same encoder to produce semantically distinct outputs. The shared signal, $\mathbf{T}_{\mathrm{shared}} = (\mathbf{T}_{\mathrm{WT}} + \mathbf{T}_{\mathrm{TC}} + \mathbf{T}_{\mathrm{ET}})/3$, is used by the bottleneck cross-attention, while the branch-specific $\mathbf{T}_{\mathrm{TC}}$ and $\mathbf{T}_{\mathrm{ET}}$ are individually routed to their respective cascade branches for text-semantic channel modulation (Sec. 2.4).

2.4 Text-Semantic Channel Modulation

While the WT boundary is largely determinable from visual appearance alone, the distinction between TC and ET sub-regions is highly semantic. Radiological texts frequently contain critical discriminative cues that directly inform which tissue subtypes are present and should be segmented. We exploit this observation by introducing text modulators that translate linguistic priors into channel-wise feature refinement signals. Each modulator follows the Squeeze-and-Excitation (SE) block paradigm [7], but we replace the conventional spatial squeeze with a textual squeeze that derives global context from the radiological reports. Specifically, the contextualized token embeddings from the LoRA-adapted BioBERT are first aggregated via global average pooling across the token dimension to obtain a sentence-level embedding. This sentence embedding is then passed through a multi-layer perceptron (MLP) to produce channel-wise modulation weights $G = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot \bar{t}))$, where $W_1 \in \mathbb{R}^{(C/r)\times 768}$ projects from the text dimension to a compressed bottleneck of dimension $C/r = 12$, and $W_2 \in \mathbb{R}^{C\times(C/r)}$ projects back to the feature dimension $C = 48$. The sigmoid output $G \in \mathbb{R}^{B\times C}$ is reshaped to $(B, C, 1, 1, 1)$ and broadcast across the spatial dimensions for element-wise multiplication with the feature map.

Two independent modulator instances are instantiated, one for TC (G_{TC}) and one for ET (G_{ET}). No text modulator is applied to the WT branch, reflecting the design principle that the tumor-vs-background distinction is primarily a visual task, while sub-region differentiation benefits from semantic guidance.

3 Experiments

3.1 Experimental Settings

Dataset. We evaluate TextSCP on the TextBraTS benchmark [17]. Each data sample comprises four co-registered MRI sequences (T1, T1ce, T2, FLAIR) with

Table 1. Comparison of TextSCP with the state-of-the-art methods on the TextBraTS dataset. † denotes the results reproduced on the same platform as ours.

Methods	Dice (%) ↑				HD95 ↓			
	ET	WT	TC	Avg.	ET	WT	TC	Avg.
3D-UNet [16]	80.4	87.3	81.6	83.1	6.11	10.51	8.93	8.17
nnU-Net [9]	82.2	87.5	82.6	84.1	<u>4.27</u>	11.90	8.52	8.23
SegResNet [5]	80.9	88.4	82.3	83.8	6.18	7.28	7.41	6.95
Swin UNETR [4]	81.0	89.5	80.8	83.8	5.95	8.23	7.03	7.07
Nestedformer [20]	82.6	89.5	80.2	84.1	5.08	10.51	8.93	8.17
TextBraTS [17]	<u>83.3</u>	<u>89.9</u>	<u>82.8</u>	<u>85.3</u>	4.58	<u>5.48</u>	**5.34**	<u>5.13</u>
TextBraTS†	82.8	89.6	82.5	84.9	5.28	8.59	6.77	6.88
TextSCP (Ours)	**85.3**	**90.7**	**85.1**	**87.0**	**3.95**	**4.98**	<u>5.51</u>	**4.81**

voxel-wise annotations for WT, TC, and ET, paired with a free-text radiological description of tumor characteristics. We follow the official train and test splits provided by the benchmark.

Evaluation Metrics. We adopt two main metrics per sub-region: **Dice** [2] measures volumetric overlap, and 95th percentile Hausdorff Distance (**HD95**) [8] measures boundary accuracy as the 95th percentile of symmetric surface distances (in mm), providing robustness to outlier errors compared to the standard Hausdorff distance. Both metrics are computed separately for WT, TC, and ET and averaged to obtain overall scores.

Implementation Details. All experiments are conducted on a single NVIDIA RTX A6000 GPU using PyTorch [14] with MONAI [1]. Input volumes are resized to 128×128 in width and height, and intensity-normalized per channel on nonzero voxels. For LoRA [6], we set rank $r = 8$ and scaling factor $\alpha = 16$ applied to the query and value projections of BioBERT [10]. The prompt length is $K = 4$ tokens per sub-region. The text-semantic channel modulators use a reduction ratio of $r = 4$. We train for 200 epochs using Sharpness-Aware Minimization (SAM) [3] with SGD [13] as the base optimizer (learning rate 0.1, momentum 0.9), combined with a linear warmup cosine annealing schedule (50 warmup epochs). The number of parameters introduced by our proposed method is approximately 0.18% of the baseline model, which is negligible.

3.2 Comparison with SOTA methods

Table 1 compares TextSCP against the SOTA methods on the TextBraTS dataset. According to this table, TextSCP achieves the highest average Dice score of 87.0%, surpassing the best previously reported result, TextBraTS, by 1.7%. The improvement is consistent across all three sub-regions. Notably, the largest gain is observed for TC (+2.3%), the sub-region most dependent on distinguishing necrotic and non-enhancing components from enhancing tissue. In terms of boundary accuracy, TextSCP achieves the best average HD95 of 4.81 mm, improving upon TextBraTS by $\approx 6.2\%$ (0.32 mm). The noticeable performance

Table 2. Ablation study of main components of our method.

Soft Cascade	Sub-Region-Aware Prompt	LoRA	Text Modulation	Dice (%) ↑	HD95 ↓
✓	×	×	×	85.9	5.67
✓	✓	×	×	86.4	5.26
✓	✓	✓	×	86.6	5.04
✓	✓	✓	✓	**87.0**	**4.81**

Table 3. Ablation of cascading approach.

Cascade Approach	Dice (%) ↑ ET	WT	TC	Avg.
WT+TC+ET	84.0	90.2	83.9	86.0
WT→TC+ET	85.0	90.5	84.8	86.7
WT→TC→ET	**85.3**	**90.7**	**85.1**	**87.0**

Table 4. Ablation of learnable tokens.

#Tokens	Dice (%) ↑ ET	WT	TC	Avg.
1	84.1	90.3	84.2	86.2
4	**85.3**	**90.7**	**85.1**	**87.0**
10	84.8	90.5	84.7	86.7

gap between TextBraTS and TextSCP highlights the advantage of moving from a shared textual representation to subregion-aware prompting. Rather than providing the same linguistic signal to all prediction targets, TextSCP adapts the text encoding to the discriminative requirements of each sub-region, leading to consistent enhancements across both overlap- and boundary-based metrics.

3.3 Ablation Study

Effect of Components. In Table 2, we progressively add each proposed component to evaluate its individual contribution. Starting from the soft cascade alone (85.9% Dice and 5.67 mm HD95), introducing sub-region-aware prompts yields a +0.5% improvement in Dice (−7.2% in HD95) by enabling BioBERT to produce specialized text representations for each tumor sub-region. Adding LoRA further improves performance to 86.6% Dice (−4.2% in HD95) by allowing the text encoder to adapt its internal representations. Finally, incorporating the text-semantic channel modulators achieves the best result of 87.0% Dice (−4.6% in HD95), as the channel-wise refinement allows the decoder to emphasize text-relevant feature channels at each cascade stage.

Effect of Cascading Approach. We compare three cascade strategies in Table 3: independent parallel heads (WT+TC+ET), partial cascading where WT guides both TC and ET simultaneously (WT→TC+ET), and our full sequential cascade (WT→TC→ET). Independent prediction performs worst (86.0%), as each sub-region must learn its boundaries without spatial priors from coarser regions. Partial cascading improves the average to 86.7%; however, the full sequential cascade achieves the best results across all sub-regions (87.0%).

Effect of Number of Learnable Tokens. We vary the number of soft prompt tokens $K \in \{1, 4, 10\}$ prepended to the text input, as demonstrated in Table 4. A single token provides limited capacity to steer BioBERT toward sub-region-specific features. K=4 achieves the best performance, providing sufficient representational capacity for sub-region specialization. Increasing to K=10 slightly

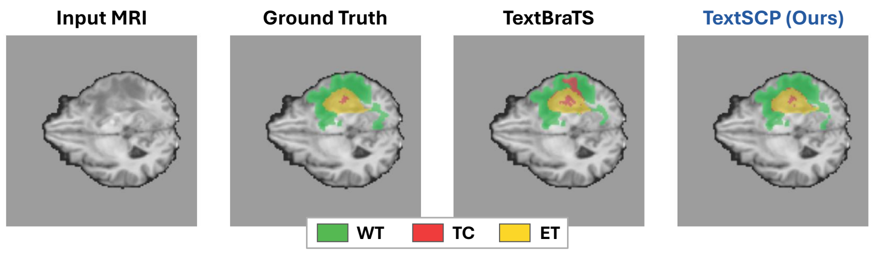

Fig. 2. Qualitative comparison with the baseline TextBraTS.

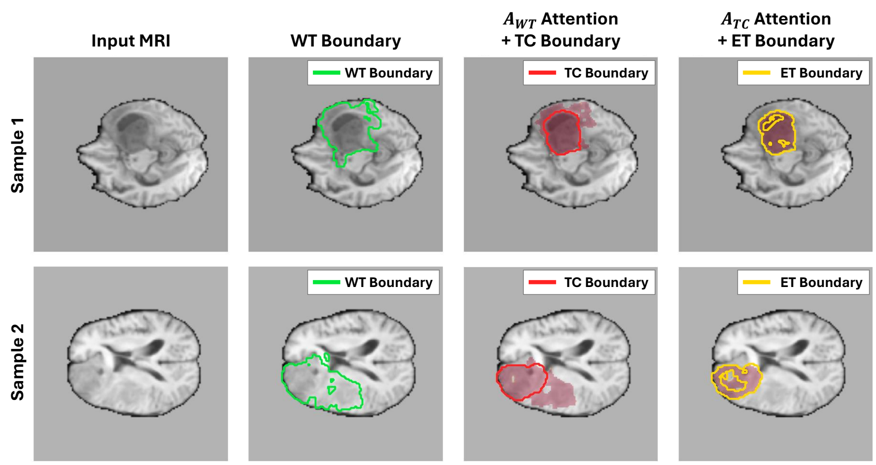

Fig. 3. Cascade attention visualization. A_{WT} concentrates within the WT region toward the TC boundary, and A_{TC} further narrows from TC toward the ET boundary.

degrades performance, consistent with the diminishing-return trend observed in the prompt-tuning literature. When the prompt length is large relative to the downstream task complexity, the additional learnable tokens tend to wash out the word-level semantics during self-attention.

3.4 Qualitative Analysis

Fig. 2 presents a qualitative comparison of segmentation results on a representative test case where TextSCP significantly outperforms our baseline. Moreover, Fig.3 provides insight into why the cascade achieves this: the WT attention map A_{WT} activates broadly across the whole tumor and concentrates toward the TC boundary, while the TC attention map A_{TC} further narrows its focus from the tumor core to tightly encompass the ET boundary. This learned coarse-to-fine spatial refinement, following the anatomical hierarchy WT$\rightarrow$TC$\rightarrow$ET, enables each cascade stage to provide a meaningful spatial prior for the subsequent finer sub-region without requiring any intermediate supervision.

4 Conclusion

We present TextSCP, a text-guided brain tumor segmentation framework that introduces a soft cascade decoder to exploit the anatomical hierarchy of tumor sub-regions and sub-region-aware prompt tuning. By prepending learnable prompts to BioBERT with lightweight LoRA adaptation, our method generates specialized text representations for each sub-region. The soft cascade progressively refines spatial attention from WT to TC and then to ET, providing each stage with an anatomical prior from its parent region. Text-semantic channel modulators further bridge the text and visual features through channel-wise refinement. Extensive experiments on the TextBraTS dataset demonstrate that TextSCP achieves SOTA performance, and we show that all components are effective.

Acknowledgments. This study was supported by the RANZCR Research Grant in 2024.

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

References

1. 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)
2. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
3. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. In: ICLR (2021)
4. 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)
5. Hsu, C., Chang, C., Chen, T.W., Tsai, H., Ma, S., Wang, W.: Brain tumor segmentation (brats) challenge short paper: Improving three-dimensional brain tumor segmentation using segresnet and hybrid boundary-dice loss. In: International MICCAI Brainlesion Workshop. pp. 334–344. Springer (2021)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
7. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks (2019)
8. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(9), 850–863 (2002)
9. 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)

10. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
11. Liu, A., Xue, R., Cao, X.R., Shen, Y., Lu, Y., Li, X., Chen, Q., Chen, J.: Medsam3: Delving into segment anything with medical concepts. arXiv preprint arXiv:2511.19046 (2025)
12. Luo, L., Tang, B., Chen, X., Han, R., Chen, T.: Vividmed: Vision language model with versatile visual grounding for medicine. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 1800–1821 (2025)
13. Nesterov, Y.: A method for solving the convex programming problem with convergence rate o (1/k2). In: Dokl akad nauk Sssr. vol. 269, p. 543 (1983)
14. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
15. Rokuss, M., Langenberg, M., Kirchhoff, Y., Isensee, F., Hamm, B., Ulrich, C., Regnery, S., Bauer, L., Katsigiannopulos, E., Norajitra, T., et al.: Voxtell: Free-text promptable universal 3d medical image segmentation. arXiv preprint arXiv:2511.11450 (2025)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI. pp. 234–241. Springer International Publishing (2015)
17. Shi, X., Jain, R.K., Li, Y., Hou, R., Cheng, J., Bai, J., Zhao, G., Lin, L., Xu, R., Chen, Y.w.: Textbrats: Text-guided volumetric brain tumor segmentation with innovative dataset development and fusion module exploration. In: MICCAI. pp. 638–648. Springer (2025)
18. Xie, Y., Wang, R., Zhuang, Y., Chen, K., Han, L., Liao, G., Hou, Y., Hou, L., Lin, J.: Tvpnet: Text-vision prompt guided segmentation for small 3d medical object. Biomedical Signal Processing and Control **111**, 108239 (2026)
19. Xin, Y., Ates, G.C., Shao, W.: Text3dsam: Text-guided 3d medical image segmentation using sam-inspired architecture. In: CVPR 2025: Foundation Models for 3D Biomedical Image Segmentation (2025)
20. Xing, Z., Yu, L., Wan, L., Han, T., Zhu, L.: Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In: MICCAI. pp. 140–150. Springer (2022)