

## Article

# NeuroSPFNet: Vision–Language Semantic Prior-Guided Frequency-Enhanced Network for Brain Tumor Segmentation in Multimodal MRI

Yantong Liu , Seong-Yoon Shin \* and Hyun-Ae Lee \*

Department of Computer Information Engineering, Kunsan National University, Gunsan 54150, Republic of Korea; lyt1994@kunsan.ac.kr

\* Correspondence: s3397220@kunsan.ac.kr (S.-Y.S.); tranhyun@gmail.com (H.-A.L.)

## Abstract

Accurate brain tumor segmentation in multimodal magnetic resonance imaging (MRI) scans is essential for quantitative neuro-oncology analysis, treatment planning, and longitudinal disease monitoring. However, automatic delineation remains challenging because glioma subregions exhibit heterogeneous appearance across T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR) sequences; weak and irregular lesion boundaries; and severe scale variation between edema, tumor core, and enhancing tumor regions. This paper presents NeuroSPFNet, a vision–language semantic prior-guided frequency-enhanced segmentation framework for tumor subregion delineation in multimodal MRI volumes. NeuroSPFNet integrates four complementary components: a Clinical Vision–Language Prior Component (CVLPC) for category-level semantic guidance; a Frequency-Enhanced Boundary Module (FEBM) for boundary-sensitive representation learning; a Multi-Scale Lesion-Dominated Fusion Module (MLDFM) for hierarchical lesion aggregation; and a Lesion-Semantic Boundary Consistency Loss (LSBCL) for jointly constraining region overlap, boundary quality, and semantic alignment. The task-specific contribution lies in the closed coupling of shared BraTS-region semantic prototypes, gated frequency responses, lesion-weighted multiscale features, and semantic-boundary supervision within one volumetric segmentation pipeline. On the Brain Tumor Segmentation (BraTS) 2021 benchmark, NeuroSPFNet achieved Dice scores of 91.94%, 88.84%, and 85.38% for whole tumor, tumor core, and enhancing tumor, respectively, with a mean Dice of 88.72% and a mean 95th-percentile Hausdorff distance (HD95) of 3.58 mm. Ablation studies show distinct effects across subregions and metrics: semantic priors primarily strengthen tumor-core and enhancing-tumor discrimination, whereas frequency and boundary constraints primarily reduce contour error. NeuroSPFNet also attains competitive performance under the current setting, with moderate computational overhead of 42.58 million parameters, 78.60 billion floating-point operations (GFLOPs), and 265 ms per volume.



Academic Editors: Nan Li and He Wei

Received: 9 June 2026

Revised: 28 July 2026

Accepted: 29 July 2026

Published: 5 August 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** brain tumor segmentation; multimodal MRI; vision–language model; clinical semantic prior; frequency enhancement; boundary-aware segmentation; multimodal medical imaging; deep learning

## 1. Introduction

Multimodal magnetic resonance imaging (MRI) is central to neuro-oncologic imaging because different MRI sequences provide complementary tissue contrast and lesion

information. In glioma assessment, T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR) sequences jointly reveal tumor morphology, edema, enhancing tissue, and surrounding anatomical context. Accurate segmentation of the whole-tumor (WT), tumor-core (TC), and enhancing-tumor (ET) regions supports quantitative analysis for treatment planning, radiotherapy target delineation, response monitoring, and imaging-biomarker extraction [1–3]. These targets preserve complementary imaging phenotypes: WT measures the complete tumor-associated abnormality, including edema; TC isolates the enhancing and non-enhancing/necrotic central tumor; and ET quantifies the contrast-enhancing viable tumor. Their separate delineation supports region-specific volumetry and reveals changes in edema, central necrosis, and enhancement that a single merged mask would obscure. Despite the clinical importance of this task, manual segmentation is time-consuming, expertise-dependent, and difficult to standardize across centers.

Automatic brain tumor segmentation from multimodal MRI scans remains challenging. First, tumor appearance is heterogeneous across patients, scanners, institutions, and imaging sequences. A tumor component may be visually prominent in one modality and weakly represented in another. Second, tumor boundaries are often ambiguous because infiltrative growth, edema, and imaging noise produce weak and irregular margins. Third, lesion subregions vary substantially in size: WT may cover a broad area, whereas ET may be small, fragmented, and heavily imbalanced relative to normal tissue. These challenges require lesion-aware representation, boundary-sensitive feature enhancement, and robust multiscale fusion.

NeuroSPFNet assigns one targeted component to each failure mode and couples the four components through a shared segmentation pathway. The Clinical Vision–Language Prior Component (CVLPC) converts MRI-oriented descriptions of WT, TC, ET, and edema into category prototypes that condition volumetric visual features, directly addressing subregions with overlapping appearance. The Frequency-Enhanced Boundary Module (FEBM) gates three-dimensional spectral responses to strengthen weak local transitions and preserve irregular lesion margins. The Multi-Scale Lesion-Dominated Fusion Module (MLDFM) reweights and aligns four decoder resolutions so that broad edema and small enhancing foci contribute within a common lesion-centered representation. The Lesion-Semantic Boundary Consistency Loss (LSBCL) jointly constrains region overlap, three-dimensional Sobel gradient agreement, and alignment between predicted lesion embeddings and the same category prototypes used by CVLPC. This allocation connects semantic ambiguity, boundary uncertainty, and scale imbalance to explicit, experimentally separable mechanisms.

Deep convolutional neural networks have become standard tools for biomedical segmentation. U-Net and its variants use encoder–decoder structures and skip connections to recover spatial detail [4–7]. Transformer-based segmentation networks further improve global context modeling through self-attention and hierarchical token representation [8–11]. Recent semantic-prior, boundary-aware, and frequency-domain studies demonstrate the value of these complementary signals [12–15]. A remaining task-level gap is the coordinated use of subregion semantics, boundary-frequency responses, and scale-dependent lesion evidence within one BraTS-oriented optimization pathway.

This study proposes NeuroSPFNet, a vision–language semantic prior-guided frequency-enhanced network for brain tumor segmentation in multimodal MRI volumes. The method uses one shared set of WT-, TC-, ET-, and edema-oriented semantic prototypes to condition intermediate visual features and to supervise class-specific output embeddings. CVLPC, FEBM, and MLDFM form the feature pathway, and LSBCL supplies region, contour, and semantic supervision at its output.

The main contributions are as follows:

- We introduce a shared-prototype semantic pathway that uses the same BraTS-oriented descriptors for intermediate visual conditioning and output-level semantic supervision.

- We couple semantic conditioning with gated volumetric frequency enhancement and lesion-weighted multiscale fusion to address subregion ambiguity, weak margins, and lesion-scale imbalance.
- We formulate LSBCL to coordinate region overlap, three-dimensional boundary consistency, and class-semantic alignment.
- We evaluate the contribution of each link through component, semantic-prior, frequency-strategy, and loss-term ablations.

## 2. Related Work

The review of the literature is organized starting with single-sequence segmentation through to multimodal and transformer-based models, followed by semantic-prior and boundary-frequency methods. This progression identifies the three linked requirements addressed by NeuroSPFNet: subregion discrimination, weak-boundary recovery, and scale-aware lesion aggregation.

### 2.1. Medical Image Segmentation

Convolutional neural network (CNN)-based biomedical segmentation methods remain highly competitive because they preserve local spatial detail and can be trained efficiently with limited data. U-Net introduced the encoder–decoder structure with skip connections [4], three-dimensional (3D) U-Net extended dense prediction to volumetric input [5], and Attention U-Net used attention gates to reduce irrelevant responses [6]. nnU-Net showed that robust preprocessing, training configuration, and postprocessing can produce strong general segmentation performance [7]. Single-sequence models establish strong spatial baselines, while their input cannot represent the distinct enhancement, edema, and tissue-relaxation information distributed across the four BraTS MRI sequences. This motivates the use of a stable U-Net-like backbone as the base architecture.

### 2.2. Brain Tumor Segmentation in Multimodal MRI

The Brain Tumor Segmentation (BraTS) benchmark has provided a widely used setting for evaluating glioma segmentation methods [1–3,16]. Typical pipelines use T1, T1ce, T2, and FLAIR sequences and report Dice and the 95th-percentile Hausdorff distance (HD95) for WT, TC, and ET. T1 provides anatomical context, T1ce highlights enhancing tissue, T2 depicts water-rich abnormalities, and FLAIR suppresses the cerebrospinal-fluid signal to emphasize edema and infiltrative change. CNN-based brain tumor segmentation has explored cascaded networks, multiscale processing, residual learning, and volumetric context [17–20]. Nevertheless, accurately segmenting small enhancing regions and fuzzy boundaries under inter-modality heterogeneity remains difficult.

### 2.3. Transformer-Based Medical Image Segmentation

Transformers model long-range dependency and have been adapted to segmentation through hybrid and hierarchical designs [21,22]. TransUNet combines CNN feature extraction with transformer encoding [8]; Swin-Unet uses shifted window attention for U-shaped segmentation [9]; and UNETR uses transformer encoders for 3D medical segmentation [10]. Recent volumetric models, including nnFormer and Swin UNETR, further strengthen hierarchical 3D context modeling [11,23]. Their progress sharpens the remaining need for explicit lesion-category conditioning and boundary-focused optimization.

### 2.4. Vision–Language Semantic Priors in Medical Imaging

Vision–language models align visual features with text descriptions and have shown broad transfer capability, exemplified by Contrastive Language–Image Pre-training (CLIP) [24]. Biomedical language and image–text models, including BioBERT, PubMedBERT, Clini-

calBERT, ConVIRT, GLoRIA, MedCLIP, and BioMedCLIP, provide useful representations for medical text or medical image–text data [25–31]. NeuroSPFNet narrows this general paradigm to four BraTS-oriented prototypes and reuses them in both feature conditioning and output-level semantic consistency.

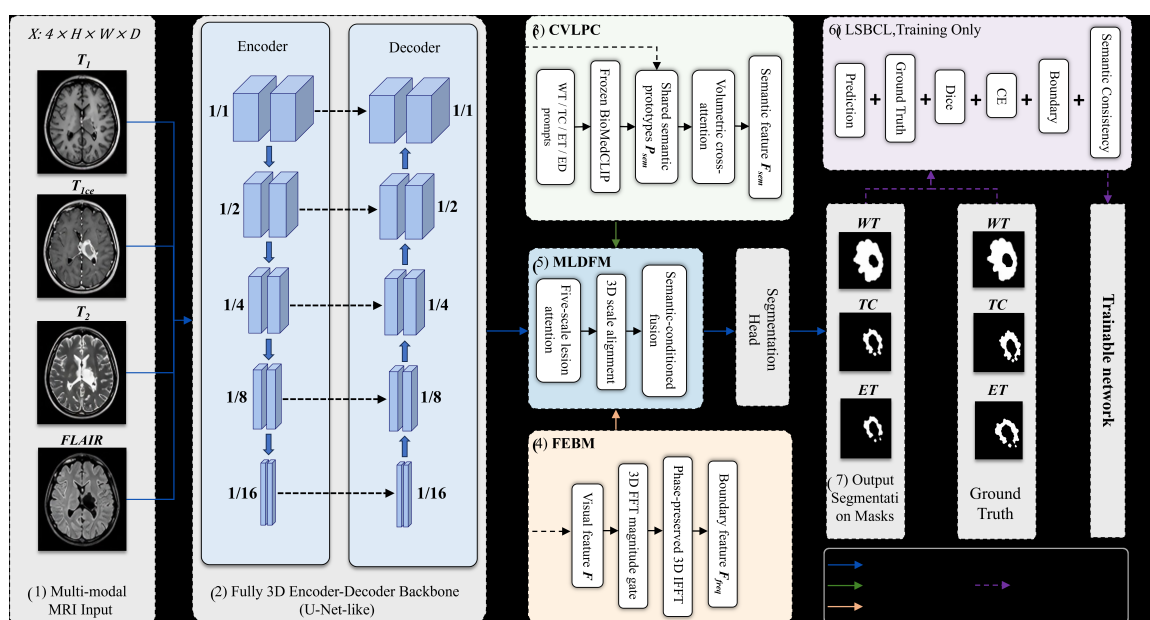
### 2.5. Frequency-Domain and Boundary-Aware Feature Enhancement

Boundary quality is critical in medical segmentation because clinical structures often have irregular contours. Boundary loss, Hausdorff-inspired loss, surface-aware objectives, and edge consistency constraints have been proposed to improve contour accuracy [14,15,32]. Frequency-domain analysis provides a complementary view of local transitions and high-frequency image detail [33–35]. NeuroSPFNet links gated three-dimensional frequency responses to Sobel gradient boundary consistency, while semantic conditioning and multiscale lesion weighting address class ambiguity and scale variation in the same pathway.

## 3. Materials and Methods

### 3.1. Overview of NeuroSPFNet

NeuroSPFNet is implemented as a three-dimensional volumetric network. Let  $X \in \mathbb{R}^{4 \times H \times W \times D}$  denote four co-registered MRI sequences, T1, T1ce, T2, and FLAIR, represented as the four input planes of the computational tensor. The term “channel” is used only for this tensor dimension; each MRI sequence measures distinct tissue contrast and physiology. A 2.5D model would stack a target slice with adjacent slices and apply two-dimensional operators, whereas the reported model uses three-dimensional convolutions,  $128 \times 128 \times 128$  subvolumes, and full-volume sliding-window inference. NeuroSPFNet augments a U-Net-like encoder–decoder with CVLPC, FEBM, MLDFM, and LSBCL, as shown in Figure 1. The overview displays all five backbone resolutions; MLDFM uses four selected decoder resolutions for the aligned fusion detailed below.



**Figure 1.** Three-dimensional NeuroSPFNet and information flow. Four co-registered MRI sequences enter the five-resolution 3D encoder–decoder. Backbone features supply the visual queries for CVLPC and the spectral input for FEBM; the resulting semantic and frequency-enhanced features join four selected decoder resolutions in MLDFM before the segmentation head predicts WT, TC, and ET. Blue, green, and orange solid arrows denote visual, semantic, and frequency-enhanced feature flow, respectively. Black dashed arrows denote encoder–decoder skip connections or auxiliary backbone feature inputs, and purple dashed arrows denote training-only LSBCL supervision.

### 3.2. Clinical Vision–Language Prior Component

CVLPC constructs semantic priors from concise clinical descriptions of tumor subregions. The prompts describe WT, TC, ET, and edema using MRI-relevant language. A frozen text encoder  $\Phi_{\text{text}}$  produces embeddings for the prompt set  $T = \{t_{\text{WT}}, t_{\text{TC}}, t_{\text{ET}}, t_{\text{ED}}\}$ . A trainable projection aligns the text embedding dimension to the visual feature dimension:

$$P_{\text{sem}} = W_p \Phi_{\text{text}}(T) + b_p, \quad (1)$$

where  $P_{\text{sem}} \in \mathbb{R}^{C_s \times d}$  contains semantic prototypes for tumor-related categories.

The exact prompts used to construct the four category prototypes were as follows:

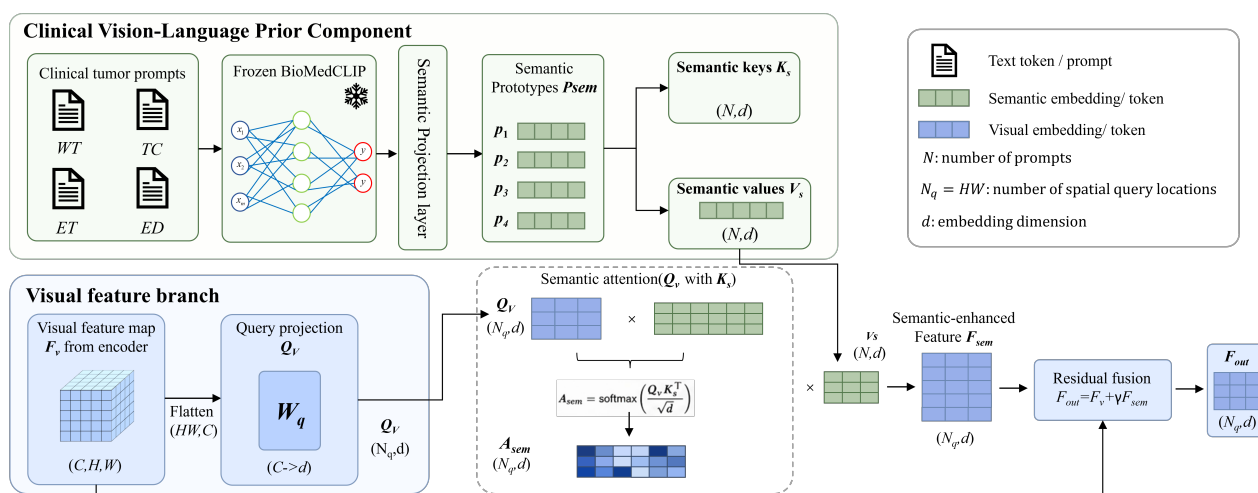
- WT: “A whole-tumor region on multimodal brain MRI, including peritumoral edema, non-enhancing or necrotic tumor core, and enhancing tumor.”
- TC: “A tumor-core region on multimodal brain MRI, including enhancing tumor and non-enhancing or necrotic core, while excluding peritumoral edema.”
- ET: “An enhancing-tumor region showing contrast enhancement on post-contrast T1-weighted MRI within the tumor core.”
- ED: “A peritumoral edema region showing hyperintensity on T2-weighted and FLAIR MRI around the tumor core.”

BioMedCLIP [31] was used as the frozen text encoder. Its parameters remained fixed, and only the semantic projection and downstream segmentation network were optimized.

For an intermediate visual feature  $F_v$ , visual queries and semantic keys/values are computed by linear projections. Specifically, the volumetric feature is rearranged from  $C \times H \times W \times D$  to  $N_q \times C$ , where  $N_q = HWD$ ; the  $HW$  layout in Figure 2 is a compact graphical projection of this volumetric tokenization. Semantic attention is defined as follows:

$$A_{\text{sem}} = \text{softmax}\left(\frac{Q_v K_s^T}{\sqrt{d}}\right), \quad F_{\text{sem}} = A_{\text{sem}} V_s, \quad F_{\text{out}} = F_v + \gamma F_{\text{sem}}. \quad (2)$$

This design introduces category-level prior guidance while keeping image evidence as the dominant segmentation signal.



**Figure 2.** Clinical Vision–Language Prior Component (CVLPC). The four WT, TC, ET, and ED prompts are encoded by frozen BioMedCLIP and projected into shared semantic prototypes. The planar token arrangement is used only to keep the schematic legible; in the implemented volumetric branch,  $F_v \in \mathbb{R}^{C \times H \times W \times D}$  is flattened into  $N_q = HWD$  visual queries. The prototypes act as semantic keys and values in cross-attention, followed by residual fusion with the 3D visual features.

### 3.3. Frequency-Enhanced Boundary Module

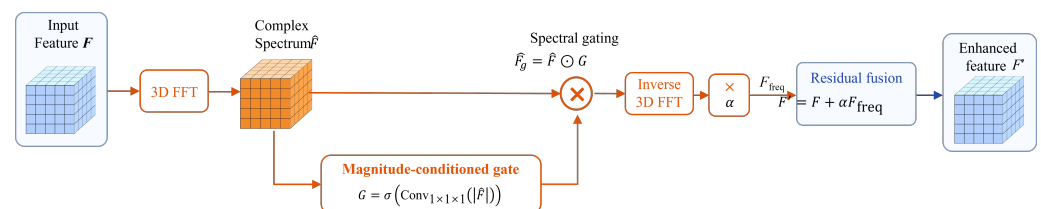
FEBM targets weak and irregular lesion boundaries by enhancing high-frequency feature responses. Given feature map  $F$ , the fast Fourier transform (FFT)-based branch computes a spectral representation, learns a frequency gate, and reconstructs an enhanced response:

$$\hat{F} = \mathcal{F}(F), \quad G = \sigma(\text{Conv}(|\hat{F}|)), \quad F_{\text{freq}} = \mathcal{F}^{-1}(\hat{F} \odot G), \quad F_{\text{FEBM}} = F + \alpha F_{\text{freq}}. \quad (3)$$

The Fourier transform is applied along the three spatial axes ( $H, W, D$ ) with orthonormal normalization. The real-valued gate is generated from the spectral magnitude by a  $1 \times 1 \times 1$  convolution and sigmoid activation, then multiplied with the complex spectrum so that the original phase is retained. The real component of the inverse transform is fused with the spatial residual through the trainable coefficient  $\alpha$ ; frequency shifting is used only for spectrum visualization. For efficient ablation, a high-frequency residual approximation can also be used:

$$F_{\text{high}} = F - \text{Up}(\text{AvgPool}(F)), \quad F_{\text{FEBM}} = F + \alpha \text{Conv}(F_{\text{high}}). \quad (4)$$

Figure 3 summarizes the frequency-domain gating and residual reconstruction pathway.



**Figure 3.** Frequency-Enhanced Boundary Module (FEBM). A 3D FFT is applied along the  $H$ ,  $W$ , and  $D$  axes. A  $1 \times 1 \times 1$  convolution predicts a real-valued gate from the spectral magnitude; this gate modulates the complex spectrum while retaining phase. The real component reconstructed by the inverse 3D FFT is scaled by the learnable coefficient  $\alpha$  and added to the unchanged spatial residual  $F$ , yielding  $F_{\text{FEBM}} = F + \alpha F_{\text{freq}}$ . Orange arrows denote frequency-domain transformation, gating, inverse transformation, and residual scaling; the blue arrow denotes the enhanced spatial-feature output.

### 3.4. Multi-Scale Lesion-Dominated Fusion Module

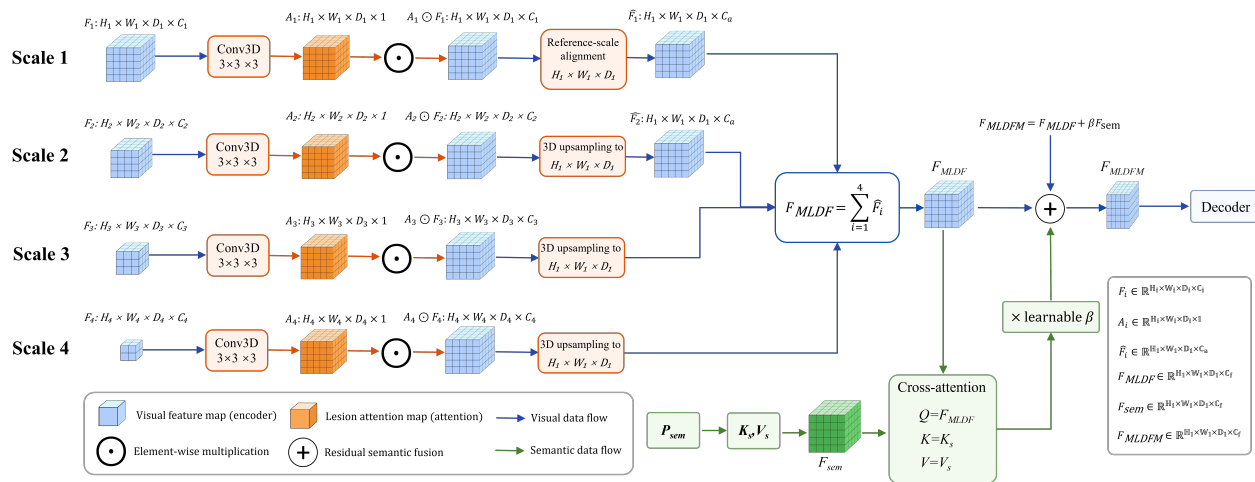
MLDFM fuses hierarchical features while emphasizing lesion-relevant regions. Four decoder resolutions are selected from the five-stage backbone. For features  $\{F_i\}_{i=1}^4$ , lesion attention maps are computed and used for three-dimensional scale-aligned fusion:

$$A_i = \sigma(\text{Conv3D}_{3 \times 3 \times 3}(F_i)), \quad \hat{F}_i = \text{Align3D}(A_i \odot F_i), \quad F_{\text{MLDF}} = \sum_{i=1}^4 \hat{F}_i. \quad (5)$$

Scale 1 defines the reference resolution ( $H_1, W_1, D_1$ ); scales 2–4 are upsampled to this spatial size, and a channel-alignment projection produces a common dimension  $C_a$ . Semantic conditioning is then applied through cross-attention:

$$F_{\text{sem}} = \text{CrossAttention}(Q = F_{\text{MLDF}}, K = K_s, V = V_s), \quad F_{\text{MLDFM}} = F_{\text{MLDF}} + \beta F_{\text{sem}}. \quad (6)$$

Figure 4 summarizes scale alignment, lesion-attention weighting, element-wise fusion, and semantic conditioning.



**Figure 4.** Multi-Scale Lesion-Dominated Fusion Module (MLDFM). Four volumetric decoder features are reweighted by  $A_i = \sigma(\text{Conv3D}_{3 \times 3 \times 3}(F_i))$ , aligned to  $(H_1, W_1, D_1)$ , and fused exclusively by element-wise summation. Cross-attention uses  $F_{MLDF}$  as the visual query and the shared CVLPC prototypes as semantic keys and values. The learnable residual coefficient  $\beta$  produces  $F_{MLDFM} = F_{MLDF} + \beta F_{sem}$  before decoder integration. Orange arrows denote lesion-attention map generation and feature reweighting, blue arrows denote multiscale visual-feature alignment and fusion, and green arrows denote semantic conditioning and its gated residual pathway.

### 3.5. Lesion-Semantic Boundary Consistency Loss

The optimization objective combines Dice loss, cross-entropy loss, boundary consistency loss, and semantic consistency loss. For each tumor-related class  $c$ , the predicted lesion embedding  $e_c^{\text{pred}}$  is aligned with the semantic prior  $p_c^{\text{sem}}$ :

$$\mathcal{L}_{\text{sem}} = 1 - \frac{1}{C} \sum_{c=1}^C \frac{e_c^{\text{pred}} \cdot p_c^{\text{sem}}}{\|e_c^{\text{pred}}\|_2 \|p_c^{\text{sem}}\|_2}. \quad (7)$$

Boundary consistency is computed from fixed three-dimensional Sobel gradients. For class  $c$ , prediction probability  $P_c$ , and binary reference  $G_c$ , the implemented term is

$$\mathcal{L}_{\text{boundary}} = \frac{1}{C} \sum_{c=1}^C \frac{1}{N} (\|S_x(P_c) - S_x(G_c)\|_1 + \|S_y(P_c) - S_y(G_c)\|_1 + \|S_z(P_c) - S_z(G_c)\|_1), \quad (8)$$

where  $S_x$ ,  $S_y$ , and  $S_z$  are fixed Sobel kernels and  $N$  is the number of voxels. The full objective is

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Dice}} + \lambda_2 \mathcal{L}_{\text{CE}} + \lambda_3 \mathcal{L}_{\text{boundary}} + \lambda_4 \mathcal{L}_{\text{sem}}. \quad (9)$$

### 3.6. Implementation Details

The implementation was developed using PyTorch version 2.1.2 (Meta Platforms, Inc., Menlo Park, CA, USA) and optimized with AdamW using an initial learning rate of  $1 \times 10^{-4}$  and cosine learning-rate decay. The five-stage 3D encoder used channel widths of 32, 64, 128, 256, and 512. Each stage contained two  $3 \times 3 \times 3$  convolutions with instance normalization and Leaky ReLU activation. Downsampling used stride-2 convolution; each decoder stage used transposed convolution, concatenation with the corresponding encoder feature, and two convolutional blocks. The prediction head produced three sigmoid region maps for WT, TC, and ET. All experiments were trained for 300 epochs with a batch size of 2 on two NVIDIA GeForce RTX 4090 graphics processing units (NVIDIA Corporation, Santa Clara, CA, USA) under deterministic random seeds. Input volumes were cropped into foreground-centered patches of size  $128 \times 128 \times 128$ , and sliding-window inference with a 50% overlap was used during evaluation. Random flipping was applied independently

along each spatial axis with probability 0.5. Random rotation was sampled uniformly from  $-15^\circ$  to  $15^\circ$ , and Gaussian noise with a standard deviation sampled from 0 to 0.1 was applied with probability 0.15. The loss weights were empirically set to  $\lambda_1 = 1.0$ ,  $\lambda_2 = 1.0$ ,  $\lambda_3 = 0.5$ , and  $\lambda_4 = 0.2$ .

### 3.7. Dataset and Preprocessing

The experiments used the 1251 labeled cases released as the BraTS 2021 training cohort [16]. Each case contains co-registered T1, T1ce, T2, and FLAIR sequences and an organizer-provided reference segmentation. We included all cases with the four required sequences and a valid reference mask; no case was excluded after the integrity check. WT was formed from BraTS labels 1, 2, and 4, TC from labels 1 and 4, and ET from label 4. Reference masks were produced and quality-controlled by the BraTS data provider under the benchmark annotation protocol. The official BraTS 2021 challenge page (<https://www.synapse.org/Synapse:syn25829067/wiki/626944>) was accessed on 28 July 2026.

A patient-level random partition assigned 1000 labeled cases to training and 251 labeled cases to held-out internal evaluation. The official 219-case BraTS 2021 validation cohort was not used for local quantitative evaluation because its reference labels are not publicly available. The partition and preprocessing protocol are summarized in Table 1. All internally generated quantitative values were computed on the 251-case held-out subset; the two explicitly labeled published-reference rows in Table 2 retain their source-reported evaluation settings. BraTS-released volumes had undergone organizer-provided co-registration, 1 mm isotropic resampling, and skull stripping. We applied no additional N4 bias-field correction; each sequence was normalized by z-score within the nonzero brain mask.

**Table 1.** BraTS 2021 partition and preprocessing protocol.

| Dataset    | MRI Sequences       | Task                                 | Training | Held-Out Evaluation | Preprocessing                                                                                                                                                                                                                                                    |
|------------|---------------------|--------------------------------------|----------|---------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| BraTS 2021 | T1, T1ce, T2, FLAIR | Brain tumor multi-class segmentation | 1000     | 251                 | Organizer-provided co-registration, 1 mm isotropic resampling and skull stripping; no additional N4 correction; nonzero-mask z-score normalization; $128 \times 128 \times 128$ foreground-centered patches; flipping, rotation, and Gaussian noise augmentation |

### 3.8. Evaluation Metrics

The Dice score and the 95th-percentile Hausdorff distance (HD95) were used for WT, TC, and ET. Dice measures volumetric overlap:

$$\text{Dice} = \frac{2TP}{2TP + FP + FN}. \quad (10)$$

HD95 is the 95th percentile of the bidirectional Euclidean surface distances between the predicted and reference masks. The distances were computed in physical space using the Neuroimaging Informatics Technology Initiative (NIfTI) voxel spacing and reported in millimeters. For every internally evaluated model and ablation configuration, the patient-level score vector contained  $n = 251$  observations. Two-sided 95% confidence intervals (CIs) for the patient-level mean were calculated using Student's  $t$  interval:

$$\text{CI}_{95\%} = \bar{x} \pm t_{0.975, n-1} \frac{s}{\sqrt{n}}, \quad (11)$$

where  $\bar{x}$  is the patient-level sample mean,  $s$  is the sample standard deviation calculated with  $n - 1$  degrees of freedom, and  $t_{0.975,250} = 1.96949776$ . Tables 2–6 report mean [95% CI]. CIs are unavailable for the two published-reference rows in Table 2 because their source studies did not provide patient-level predictions or score vectors. The intervals quantify uncertainty in the reported means; no pairwise hypothesis test is claimed.

**Table 2.** Comparison with models reproduced under the unified study protocol and published BraTS 2021 results. Cells report mean [95% confidence interval]. HD95 values are in millimeters. Confidence intervals are reported for all models evaluated under the unified 251-patient protocol. N/A indicates that confidence intervals are unavailable for the two published-reference rows because patient-level predictions and score vectors were not provided by the source studies. Boldface is limited to the reproduced-model block and denotes the best point estimate within that block.

| Method                           | Backbone    | WT Dice                        | TC Dice                        | ET Dice                        | Mean Dice                      | WT HD95                     | TC HD95                     | ET HD95                     | Mean HD95                   |
|----------------------------------|-------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
| <i>CNN-based methods</i>         |             |                                |                                |                                |                                |                             |                             |                             |                             |
| U-Net                            | CNN         | 89.24<br>[87.81, 90.67]        | 83.56<br>[81.07, 86.05]        | 78.45<br>[74.79, 82.11]        | 83.75<br>[81.64, 85.86]        | 5.84<br>[4.70, 6.98]        | 7.32<br>[5.46, 9.18]        | 4.95<br>[2.71, 7.19]        | 6.04<br>[4.67, 7.41]        |
| Attention U-Net                  | CNN         | 89.85<br>[88.47, 91.23]        | 84.42<br>[82.03, 86.81]        | 79.18<br>[75.64, 82.72]        | 84.48<br>[82.44, 86.52]        | 5.52<br>[4.41, 6.63]        | 6.98<br>[5.20, 8.76]        | 4.62<br>[2.48, 6.76]        | 5.71<br>[4.39, 7.03]        |
| 3D U-Net                         | CNN         | 90.52<br>[89.20, 91.84]        | 85.80<br>[83.56, 88.04]        | 81.35<br>[78.12, 84.58]        | 85.89<br>[84.00, 87.78]        | 5.10<br>[4.07, 6.13]        | 6.25<br>[4.61, 7.89]        | 3.84<br>[1.87, 5.81]        | 5.06<br>[3.85, 6.27]        |
| nnU-Net-style 3D U-Net           | CNN         | 91.65<br>[90.43, 92.87]        | 88.42<br>[86.42, 90.42]        | 84.76<br>[81.88, 87.64]        | 88.28<br>[86.65, 89.91]        | 4.25<br>[3.32, 5.18]        | 4.85<br>[3.36, 6.34]        | 2.92<br>[1.17, 4.67]        | 4.01<br>[2.92, 5.10]        |
| <i>Transformer-based methods</i> |             |                                |                                |                                |                                |                             |                             |                             |                             |
| TransUNet                        | Hybrid      | 90.38<br>[89.05, 91.71]        | 85.95<br>[83.74, 88.16]        | 81.62<br>[78.43, 84.81]        | 85.98<br>[84.12, 87.84]        | 5.15<br>[4.10, 6.20]        | 6.18<br>[4.56, 7.80]        | 3.75<br>[1.82, 5.68]        | 5.03<br>[3.85, 6.21]        |
| Swin-Unet                        | Pure TF     | 90.95<br>[89.67, 92.23]        | 86.48<br>[84.34, 88.62]        | 82.50<br>[79.42, 85.58]        | 86.64<br>[84.84, 88.44]        | 4.88<br>[3.89, 5.87]        | 5.74<br>[4.20, 7.28]        | 3.42<br>[1.58, 5.26]        | 4.68<br>[3.56, 5.80]        |
| UNETR                            | Hybrid      | 91.50<br>[90.27, 92.73]        | 88.65<br>[86.64, 90.66]        | 84.20<br>[81.28, 87.12]        | 88.12<br>[86.48, 89.76]        | 4.12<br>[3.18, 5.06]        | <b>4.35</b><br>[2.85, 5.85] | <b>2.40</b><br>[0.63, 4.17] | 3.62<br>[2.51, 4.73]        |
| NeuroSPFNet (ours)               | Ours        | <b>91.94</b><br>[90.70, 93.18] | <b>88.84</b><br>[86.85, 90.83] | <b>85.38</b><br>[82.52, 88.24] | <b>88.72</b><br>[87.10, 90.34] | <b>3.78</b><br>[2.85, 4.71] | 4.48<br>[2.99, 5.97]        | 2.48<br>[0.74, 4.22]        | <b>3.58</b><br>[2.46, 4.70] |
| nnFormer [11,36]                 | Transformer | 91.46<br>[N/A]                 | 87.42<br>[N/A]                 | 82.22<br>[N/A]                 | 87.03<br>[N/A]                 | 10.15<br>[N/A]              | 9.59<br>[N/A]               | 16.78<br>[N/A]              | 12.17<br>[N/A]              |
| Swin UNETR [23]                  | Transformer | 92.70<br>[N/A]                 | 87.60<br>[N/A]                 | 85.30<br>[N/A]                 | 88.53<br>[N/A]                 | 4.74<br>[N/A]               | 15.31<br>[N/A]              | 16.33<br>[N/A]              | 12.13<br>[N/A]              |

### 3.9. Experimental Protocol

All reproduced models used the same 1000/251 patient partition and evaluation script. The nnU-Net-style comparator was trained under this unified protocol and is identified as such in Table 2; it does not represent the complete native self-configuring nnU-Net pipeline. Published nnFormer and Swin UNETR results are reported separately as BraTS 2021 reference values because their evaluation partitions and training recipes differ from the present internal protocol. Table 3 adds CVLPC, FEBM, MLDFM, and LSBCL sequentially to the same baseline. Table 4 changes only the semantic prior source; Table 5 changes only the frequency-enhancement strategy; and Table 6 changes only the active loss terms; all remaining settings are held fixed.

**Table 3.** Component-ablation study of NeuroSPFNet. Score cells report mean [95% confidence interval] over the unified 251-patient evaluation set. HD95 values are in millimeters. Boldface denotes the best point estimate in this table.

| Configuration     | Primary Effect        | CVLPC | FEBM | MLDFM | LSBCL | WT Dice                        | TC Dice                        | ET Dice                        | Mean Dice                      | $\Delta$ Dice | Mean HD95                   | $\Delta$ HD95 |
|-------------------|-----------------------|-------|------|-------|-------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|---------------|-----------------------------|---------------|
| Baseline          | Reference model       | –     | –    | –     | –     | 90.12<br>[88.75, 91.49]        | 84.85<br>[82.46, 87.24]        | 80.52<br>[76.94, 84.10]        | 85.16<br>[83.30, 87.02]        | –             | 5.24<br>[3.95, 6.53]        | –             |
| +CVLPC            | TC/ET discrimination  | ✓     | –    | –     | –     | 90.65<br>[89.34, 91.96]        | 86.72<br>[84.55, 88.89]        | 83.15<br>[79.98, 86.32]        | 86.84<br>[85.10, 88.58]        | +1.68         | 4.95<br>[3.74, 6.16]        | –0.29         |
| +CVLPC+FEBM       | Boundary quality      | ✓     | ✓    | –     | –     | 90.58<br>[89.30, 91.86]        | 87.25<br>[85.14, 89.36]        | 84.31<br>[81.26, 87.36]        | 87.38<br>[85.68, 89.08]        | +0.54         | 4.28<br>[3.14, 5.42]        | –0.67         |
| +CVLPC+FEBM+MLDFM | WT/TC scale fusion    | ✓     | ✓    | ✓     | –     | 91.30<br>[90.04, 92.56]        | 87.95<br>[85.90, 90.00]        | 84.78<br>[81.82, 87.74]        | 88.01<br>[86.36, 89.66]        | +0.63         | 4.16<br>[3.03, 5.29]        | –0.12         |
| Full NeuroSPFNet  | Dice-boundary balance | ✓     | ✓    | ✓     | ✓     | <b>91.94</b><br>[90.70, 93.18] | <b>88.84</b><br>[86.85, 90.83] | <b>85.38</b><br>[82.52, 88.24] | <b>88.72</b><br>[87.10, 90.34] | +0.71         | <b>3.58</b><br>[2.46, 4.70] | –0.58         |

**Table 4.** Semantic prior source ablation for CVLPC. Score cells report mean [95% confidence interval] over the unified 251-patient evaluation set. HD95 values are in millimeters. Boldface denotes the best point estimate in this table.

| Prior Source      | Frozen | WT Dice                               | TC Dice                               | ET Dice                               | Mean Dice                             | Mean HD95                          |
|-------------------|--------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|------------------------------------|
| None              | –      | 90.12<br>[88.75, 91.49]               | 84.85<br>[82.46, 87.24]               | 80.52<br>[76.94, 84.10]               | 85.16<br>[83.30, 87.02]               | 5.24<br>[3.95, 6.53]               |
| Random embedding  | No     | 90.15<br>[88.79, 91.51]               | 84.92<br>[82.54, 87.30]               | 80.59<br>[77.02, 84.16]               | 85.22<br>[83.37, 87.07]               | 5.23<br>[3.95, 6.51]               |
| Learned embedding | No     | 90.32<br>[88.98, 91.66]               | 85.46<br>[83.13, 87.79]               | 81.41<br>[77.94, 84.88]               | 85.73<br>[83.93, 87.53]               | 5.10<br>[3.84, 6.36]               |
| CLIP text encoder | Yes    | 90.50<br>[89.17, 91.83]               | 85.88<br>[83.59, 88.17]               | 80.60<br>[77.17, 84.03]               | 85.66<br>[83.84, 87.48]               | 5.18<br>[3.91, 6.45]               |
| BioClinicalBERT   | Yes    | 90.54<br>[89.22, 91.86]               | 86.26<br>[84.02, 88.50]               | 82.13<br>[78.80, 85.46]               | 86.31<br>[84.55, 88.07]               | 5.01<br>[3.78, 6.24]               |
| BioMedCLIP        | Yes    | <b>90.65</b><br><b>[89.34, 91.96]</b> | <b>86.72</b><br><b>[84.55, 88.89]</b> | <b>83.15</b><br><b>[79.98, 86.32]</b> | <b>86.84</b><br><b>[85.10, 88.58]</b> | <b>4.95</b><br><b>[3.74, 6.16]</b> |

**Table 5.** Frequency enhancement strategy ablation. Score cells report mean [95% confidence interval] over the unified 251-patient evaluation set. HD95 values are in millimeters. Boldface denotes the best point estimate in this table.

| Strategy                | Mean Dice                             | Mean HD95                          | Parameter Increase      |
|-------------------------|---------------------------------------|------------------------------------|-------------------------|
| None                    | 86.84<br>[85.10, 88.58]               | 4.95<br>[3.74, 6.16]               | 0                       |
| Sobel edge enhancement  | 87.05<br>[85.34, 88.76]               | 4.40<br>[3.24, 5.56]               | Minimal                 |
| High-frequency residual | 87.47<br>[85.78, 89.16]               | 4.52<br>[3.35, 5.69]               | Small                   |
| FFT frequency gating    | <b>87.56</b><br><b>[85.88, 89.24]</b> | <b>4.15</b><br><b>[3.03, 5.27]</b> | Moderate but acceptable |

**Table 6.** Loss function ablation for LSBCL. Score cells report mean [95% confidence interval] over the unified 251-patient evaluation set. HD95 values are in millimeters. Boldface denotes the best point estimate in this table.

| Loss Setting                         | WT Dice                               | TC Dice                               | ET Dice                               | Mean Dice                             | Mean HD95                          |
|--------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|------------------------------------|
| Dice + cross-entropy                 | 91.30<br>[90.04, 92.56]               | 87.95<br>[85.90, 90.00]               | 84.78<br>[81.82, 87.74]               | 88.01<br>[86.36, 89.66]               | 4.16<br>[3.03, 5.29]               |
| Dice + cross-entropy + boundary loss | 91.62<br>[90.37, 92.87]               | 88.12<br>[86.09, 90.15]               | 84.80<br>[81.86, 87.74]               | 88.18<br>[86.54, 89.82]               | 3.70<br>[2.62, 4.78]               |
| Dice + cross-entropy + semantic loss | 91.35<br>[90.10, 92.60]               | 88.72<br>[86.72, 90.72]               | 85.31<br>[82.43, 88.19]               | 88.46<br>[86.83, 90.09]               | 3.92<br>[2.83, 5.01]               |
| Full LSBCL                           | <b>91.94</b><br><b>[90.70, 93.18]</b> | <b>88.84</b><br><b>[86.85, 90.83]</b> | <b>85.38</b><br><b>[82.52, 88.24]</b> | <b>88.72</b><br><b>[87.10, 90.34]</b> | <b>3.58</b><br><b>[2.46, 4.70]</b> |

## 4. Experiments and Results

### 4.1. Main Comparison

Table 2 separates models reproduced on the 251-case held-out subset from published BraTS 2021 results reported under source-specific protocols. Within the reproduced block, NeuroSPFNet yielded a mean Dice of 88.72% [87.10%, 90.34%] and a mean HD95 of 3.58 mm [2.46 mm, 4.70 mm]. The widest class-specific interval occurred for ET Dice, 85.38% [82.52%, 88.24%], reflecting greater interpatient variability for the smallest tumor subregion. The nnU-Net-style comparator used the unified study protocol and is identified separately from the complete native self-configuring nnU-Net pipeline. Published nnFormer and Swin UNETR values position the method against recent volumetric transformers under source-specific evaluation settings. Their CIs remain unavailable because the source studies did

not release patient-level predictions or score vectors. Within the unified block, overlapping CIs identify several close comparisons and support interpreting the reported ordering as point-estimate evidence rather than a formal pairwise significance result.

#### 4.2. Ablation Study

The ablation protocol isolates the contribution of each proposed component. Table 3 evaluates CVLPC, FEBM, MLDFM, and LSBCL. Table 4 compares the semantic prior sources. Table 5 evaluates the frequency strategies. Table 6 evaluates the loss components. Each ablation changes only the factor named in the corresponding table while retaining the same patient partition, optimizer, training duration, inference procedure, and evaluation script.

The component-ablation results in Table 3 show complementary numerical patterns across the four modules. Relative to the baseline mean Dice of 85.16% [83.30%, 87.02%], the CVLPC configuration reached 86.84% [85.10%, 88.58%], with the largest point-estimate changes in TC and ET. Adding FEBM increased the mean Dice to 87.38% [85.68%, 89.08%] and reduced the mean HD95 to 4.28 mm [3.14 mm, 5.42 mm], while WT Dice remained stable. MLDFM increased the mean Dice to 88.01% [86.36%, 89.66%]. The full configuration reached 88.72% [87.10%, 90.34%] mean Dice and 3.58 mm [2.46 mm, 4.70 mm] mean HD95. The intervals expose substantial patient-level variability, while the sequential point estimates provide a consistent Dice increase and HD95 reduction across the complete component pathway.

Table 4 shows distinct patterns across prior sources. Random embeddings produced a mean Dice of 85.22% [83.37%, 87.07%], close to the no-prior configuration at 85.16% [83.30%, 87.02%]. Learned embeddings and the generic CLIP text encoder produced intermediate point estimates. BioClinicalBERT and BioMedCLIP produced the strongest numerical results in this ablation, with BioMedCLIP reaching 86.84% [85.10%, 88.58%] mean Dice and 4.95 mm [3.74 mm, 6.16 mm] mean HD95. The concentration of the point-estimate gains in TC and ET is consistent with the proposed role of clinically aligned priors in tumor-subregion discrimination.

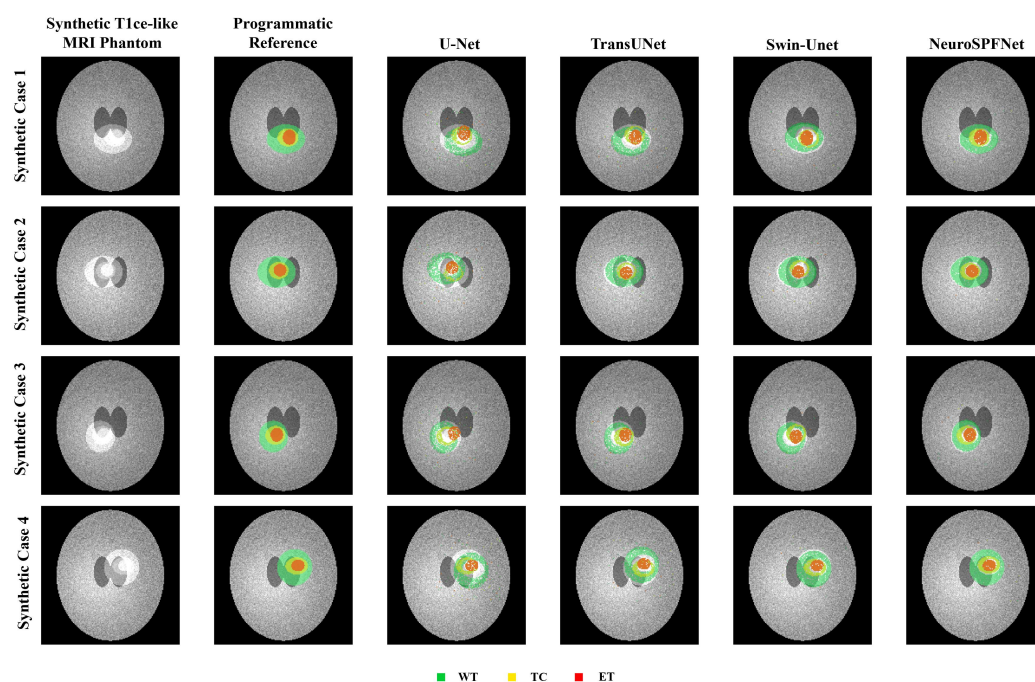
The frequency enhancement ablation in Table 5 places fast Fourier transform (FFT) frequency gating at the strongest observed Dice and HD95 operating point. Sobel edge enhancement yielded 87.05% [85.34%, 88.76%] mean Dice and 4.40 mm [3.24 mm, 5.56 mm] mean HD95, while the high-frequency residual strategy reached 87.47% [85.78%, 89.16%] and 4.52 mm [3.35 mm, 5.69 mm]. Relative to the non-frequency-enhanced setting, FFT gating changed mean Dice from 86.84% [85.10%, 88.58%] to 87.56% [85.88%, 89.24%] and mean HD95 from 4.95 mm [3.74 mm, 6.16 mm] to 4.15 mm [3.03 mm, 5.27 mm].

The loss ablation in Table 6 shows distinct metric responses across the loss terms. Boundary loss reduced mean HD95 from 4.16 mm [3.03 mm, 5.29 mm] to 3.70 mm [2.62 mm, 4.78 mm]. Semantic loss produced the strongest ET Dice point estimate among the single-added-loss settings, reaching 85.31% [82.43%, 88.19%], with a mean Dice of 88.46% [86.83%, 90.09%]. The full LSBCL configuration reached the highest observed mean Dice, 88.72% [87.10%, 90.34%], and the lowest observed mean HD95, 3.58 mm [2.46 mm, 4.70 mm]. The overlapping intervals quantify the scale of patient-level uncertainty, while the metric-specific point-estimate pattern aligns with the designated role of each loss term.

#### 4.3. Visualization Analysis

Figures 5–7 present controlled qualitative stress tests on programmatically generated two-dimensional MRI-like phantoms. The phantoms contain nested WT, TC, and ET regions with controlled intensity heterogeneity, boundary irregularity, spatial displacement, and Gaussian noise. NeuroSPFNet and all comparators were trained exclusively on the 1000-patient BraTS training partition. The phantoms were generated after training and were

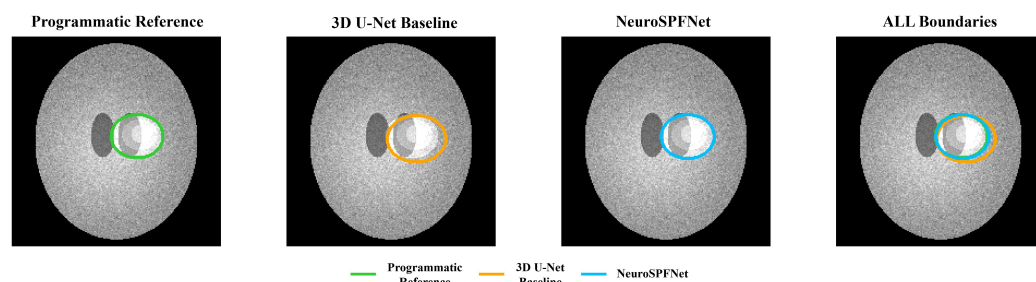
excluded from parameter optimization, model selection, quantitative evaluation, and every BraTS result in Tables 2–6. Their simplified geometry supplies exact programmatic masks and controlled perturbations, allowing segmentation leakage, contour displacement, and class-specific response localization to be inspected without anatomical and acquisition factors changing simultaneously. The resulting panels provide mechanism-oriented stress tests that complement the patient-level quantitative evaluation; they carry no estimate of in vivo performance.



**Figure 5.** Controlled segmentation stress test on synthetic MRI-like phantoms. Each row shows a synthetic T1ce-like slice, the programmatic reference mask, and predictions from U-Net, TransUNet, Swin-Unet, and NeuroSPFNet. These phantoms were used only for qualitative visualization and were excluded from every BraTS quantitative result.

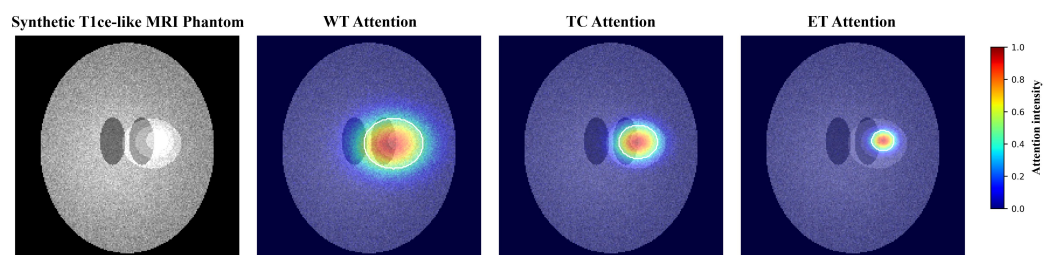
Figure 5 tests whether each network preserves the nested WT–TC–ET organization across four changes in lesion position, surrounding intensity, and noise. U-Net produces the most fragmented masks and scattered false-positive voxels, while TransUNet and Swin-Unet recover the principal lesion geometry but retain local leakage and uneven subregion boundaries. NeuroSPFNet yields the closest visual correspondence to the programmatic masks across the four cases, particularly for the compact ET center and the continuity of the surrounding TC and WT regions. This pattern is consistent with the quantitative ablations: category prototypes support nested subregion discrimination, and lesion-weighted multiscale fusion stabilizes regions with markedly different spatial extent.

Figure 6 isolates outer-contour behavior in a case with an irregular, low-contrast margin. The 3D U-Net contour shows a visible outward displacement along the inferior and right margins. The NeuroSPFNet contour follows the programmatic boundary more closely over most of the circumference and preserves the asymmetric lesion shape. The localized reduction in contour deviation provides a visual mechanism-level counterpart to the lower HD95 observed after FEBM and boundary-consistency supervision in Tables 5 and 6.



**Figure 6.** Boundary stress test on a synthetic T1ce-like MRI phantom. Green, orange, and cyan contours denote the programmatic reference, 3D U-Net baseline prediction, and NeuroSPFNet prediction, respectively. The panel isolates contour behavior under controlled contrast and noise and does not show a patient scan.

Figure 7 examines whether the shared semantic prototypes produce spatially ordered class responses. WT attention covers the full abnormal region, TC attention contracts toward the central tumor compartment, and ET attention forms the most compact response around the enhancing center. The progressive reduction in response extent follows the programmed containment relationship  $ET \subset TC \subset WT$ , while activation outside the lesion remains low. This class-dependent localization supports the proposed role of CVLPC and the semantic term in separating nested tumor targets with overlapping image evidence.



**Figure 7.** Class-specific attention on a synthetic T1ce-like MRI phantom. Heat maps show normalized WT, TC, and ET responses on controlled nested-lesion geometry; white contours denote the programmatic class boundaries. The phantom was used exclusively for qualitative inspection.

#### 4.4. Computational Efficiency

Table 7 reports parameter count, billions of floating-point operations (GFLOPs), inference time, and mean Dice. NeuroSPFNet achieved the highest mean Dice in this comparison with moderate overhead. Specifically, the model required 42.58 M parameters, 78.60 GFLOPs, and 265 ms per volume, making it slower and larger than the nnU-Net-style 3D U-Net comparator while still being more compact than TransUNet and UNETR.

**Table 7.** Computational efficiency comparison. Boldface denotes the best mean Dice point estimate in this table.

| Method                 | Params (M) | GFLOPs | Inference Time (ms/volume) | Mean Dice    |
|------------------------|------------|--------|----------------------------|--------------|
| U-Net                  | 31.04      | 65.20  | 185                        | 83.75        |
| Swin-Unet              | 27.15      | 58.40  | 210                        | 86.64        |
| TransUNet              | 105.32     | 112.50 | 380                        | 85.98        |
| nnU-Net-style 3D U-Net | 30.85      | 72.80  | 245                        | 88.28        |
| UNETR                  | 92.64      | 95.20  | 340                        | 88.12        |
| NeuroSPFNet (Ours)     | 42.58      | 78.60  | 265                        | <b>88.72</b> |

## 5. Discussion

CVLPC introduces clinical semantic priors as category-level guidance for tumor subregions. The semantic-prior point estimates place random embeddings close to the no-prior setting, learned embeddings above that reference, and the biomedical text–image priors at the strongest observed Dice–HD95 operating points. The numerical pattern is concentrated in TC and ET, where visual ambiguity and class imbalance are common. The shared prototypes influence both intermediate feature selection and output embedding consistency, which provides a mechanism-level explanation for the observed concentration in subregion discrimination. The semantic branch supplies category-level descriptors for computer-assisted segmentation and does not perform diagnostic reasoning.

FEBM is motivated by the weak and irregular boundaries observed in MRI tumor segmentation. In the frequency ablation, FFT-based gating changed the mean HD95 point estimate from 4.95 mm to 4.15 mm and the mean Dice point estimate from 86.84% to 87.56%. Sobel filtering produced its clearest numerical change in boundary distance, and high-frequency residual enhancement produced a Dice point estimate close to FFT with a higher HD95 point estimate. The alignment between these descriptive results and FEBM's boundary-targeted construction supports continued evaluation of gated high-frequency responses for lesion-margin delineation. The larger change in HD95 than in Dice is consistent with FEBM's operating target: the gate emphasizes spectral responses associated with local transitions, while the Sobel gradient term suppresses isolated contour deviations.

MLDFM addresses the scale gap between large edema regions and small enhancing components. In the progressive ablation, introducing MLDFM after CVLPC and FEBM changed the mean Dice point estimate from 87.38% to 88.01% and the mean HD95 point estimate from 4.28 mm to 4.16 mm. The stronger numerical change in overlap than in boundary distance is consistent with a scale-sensitive regional representation across heterogeneous lesion sizes.

The observed numerical pattern across WT, TC, and ET follows the structure of the task. WT is larger and strongly expressed on T2/FLAIR, leaving less headroom for semantic disambiguation. TC requires separation of edema from central tumor tissue, while ET is small, contrast-dependent, and strongly imbalanced. CVLPC produces its largest observed changes in TC and ET, FEBM produces its clearest change in boundary distance, and MLDFM supplies the scale context needed to connect small enhancing regions with the broader lesion hierarchy.

NeuroSPFNet provides automated WT, TC, and ET measurements for research-oriented quantitative analysis. Its present evidence establishes benchmark-level segmentation performance; clinical workflow benefit requires dedicated external and prospective evaluation.

This study has five principal limitations. Evaluation is restricted to one public benchmark and one internal 251-patient partition; the reported 95% CIs quantify patient-level uncertainty within this partition, and several close model comparisons have overlapping intervals. The four semantic prompts are manually specified and remain sensitive to wording and text-encoder choice. The controlled synthetic phantoms inspect mechanism behavior but do not represent the diversity of *in vivo* MRI. The semantic and spectral branches add computation relative to the base network. External multi-center validation, paired model-comparison analysis on prospectively retained score vectors, missing-sequence testing, prompt robustness, uncertainty calibration, and workflow-level evaluation define the next validation stage.

## 6. Conclusions

This work presents NeuroSPFNet, a vision–language semantic prior-guided frequency-enhanced framework for brain tumor segmentation in multimodal MRI volumes. By combining CVLPC, FEBM, MLDFM, and LSBCL, the proposed method targets tumor-subregion discrimination, boundary delineation, and multiscale lesion representation. On the 251-case held-out BraTS 2021 subset, NeuroSPFNet yielded Dice scores of 91.94% [90.70%, 93.18%], 88.84% [86.85%, 90.83%], and 85.38% [82.52%, 88.24%] for WT, TC, and ET, respectively, with a mean Dice of 88.72% [87.10%, 90.34%] and a mean HD95 of 3.58 mm [2.46 mm, 4.70 mm]. The component, prior-source, frequency-strategy, and loss-term ablations produce complementary patterns across subregion discrimination and contour accuracy, with 95% CIs reporting the patient-level uncertainty of every internally evaluated score. NeuroSPFNet also presents moderate computational overhead under the current setting. The results position the coupling of shared semantic prototypes, frequency-enhanced boundary modeling, and multiscale lesion fusion as a testable benchmark contribution. External multi-center validation and paired model-comparison analysis define the next evaluation stage.

**Author Contributions:** Conceptualization, Y.L. and S.-Y.S.; methodology, Y.L., S.-Y.S., and H.-A.L.; software, Y.L.; validation, Y.L., H.-A.L., and S.-Y.S.; formal analysis, Y.L.; investigation, Y.L.; resources, H.-A.L. and S.-Y.S.; data curation, Y.L.; writing, original draft preparation, Y.L.; writing—review and editing, Y.L., H.-A.L., and S.-Y.S.; visualization, Y.L.; supervision, H.-A.L. and S.-Y.S.; project administration, H.-A.L. and S.-Y.S.; funding acquisition, H.-A.L. and S.-Y.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in Software, supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) in 2026 (No. 2023-0-00065).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The 1251 labeled BraTS 2021 cases used for model development are available through the official BraTS 2021 Synapse portal (<https://www.synapse.org/Synapse:syn25829067/wiki/626944>, accessed on 28 July 2026), subject to the benchmark terms and data use agreement. The 219 label-withheld challenge cases were not used to calculate the locally reported metrics.

**Acknowledgments:** The authors acknowledge the organizers of the BraTS benchmark and the open-source medical image analysis community.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

|       |                                          |
|-------|------------------------------------------|
| MRI   | Magnetic resonance imaging               |
| BraTS | Brain Tumor Segmentation benchmark       |
| VLM   | Vision-language model                    |
| CNN   | Convolutional neural network             |
| FFT   | Fast Fourier transform                   |
| CE    | Cross-entropy                            |
| PLM   | Pretrained language model                |
| WT    | Whole tumor                              |
| TC    | Tumor core                               |
| ET    | Enhancing tumor                          |
| ED    | Edema                                    |
| CVLPC | Clinical Vision–Language Prior Component |

|       |                                            |
|-------|--------------------------------------------|
| FEBM  | Frequency-Enhanced Boundary Module         |
| MLDFM | Multi-Scale Lesion-Dominated Fusion Module |
| LSBCL | Lesion-Semantic Boundary Consistency Loss  |
| HD95  | 95th percentile Hausdorff distance         |

## References

- Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34*, 1993–2024. <https://doi.org/10.1109/TMI.2014.2377694>.
- Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.S.; Freymann, J.B.; Farahani, K.; Davatzikos, C. Advancing the Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features. *Sci. Data* **2017**, *4*, 170117. <https://doi.org/10.1038/sdata.2017.117>.
- Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R.T.; Berger, C.; Ha, S.M.; Rozycki, M.; et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. *arXiv* **2018**, arXiv:1811.02629.
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241. [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28).
- Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2016; pp. 424–432. [https://doi.org/10.1007/978-3-319-46723-8\\_49](https://doi.org/10.1007/978-3-319-46723-8_49).
- Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. In *Proceedings of the Medical Imaging with Deep Learning*, Amsterdam, The Netherlands, 4–6 July 2018.
- 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. *Nat. Methods* **2021**, *18*, 203–211. <https://doi.org/10.1038/s41592-020-01008-z>.
- 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* **2021**, arXiv:2102.04306.
- 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 Workshops*; Springer: Cham, Switzerland, 2022; pp. 205–218.
- Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. UNETR: Transformers for 3D Medical Image Segmentation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*; IEEE: Piscataway, NJ, USA, 2022; pp. 574–584.
- Zhou, H.Y.; Guo, J.; Zhang, Y.; Han, X.; Yu, L.; Wang, L.; Yu, Y. nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer. *IEEE Trans. Image Process.* **2023**, *32*, 4036–4045. <https://doi.org/10.1109/TIP.2023.3293771>.
- Liu, J.; Zhang, Y.; Chen, J.; Xiao, J.; Lu, Y.; A Landman, B.; Yuan, Y.; Yuille, A.; Tang, Y.; Zhou, Z. CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection. In *IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2023.
- Luo, H.; Bao, J.; Wu, Y.; He, X.; Li, T. CLIP4Seg: Image Semantic Segmentation via CLIP. *arXiv* **2022**, arXiv:2112.10003.
- Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, E.; Dolz, J.; Ayed, I.B. Boundary Loss for Highly Unbalanced Segmentation. *Med. Image Anal.* **2021**, *67*, 101851. <https://doi.org/10.1016/j.media.2020.101851>.
- Ma, J. Loss Odyssey in Medical Image Segmentation. *Med. Image Anal.* **2021**, *71*, 102035. <https://doi.org/10.1016/j.media.2021.102035>.
- Baid, U.; Ghodasara, S.; Mohan, S.; Bilello, M.; Calabrese, E.; Colak, E.; Farahani, K.; Kalpathy-Cramer, J.; Kitamura, F.C.; Pati, S.; et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. *arXiv* **2021**, arXiv:2107.02314.
- Havaei, M.; Davy, A.; Warde-Farley, D.; Biard, A.; Courville, A.; Bengio, Y.; Pal, C.; Jodoin, P.M.; Larochelle, H. Brain Tumor Segmentation with Deep Neural Networks. *Med. Image Anal.* **2017**, *35*, 18–31. <https://doi.org/10.1016/j.media.2016.05.004>.
- Kamnitsas, K.; Ledig, C.; Newcombe, V.F.J.; Simpson, J.P.; Kane, A.D.; Menon, D.K.; Rueckert, D.; Glocker, B. Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. *Med. Image Anal.* **2017**, *36*, 61–78. <https://doi.org/10.1016/j.media.2016.10.004>.
- Wang, G.; Li, W.; Ourselin, S.; Vercauteren, T. Automatic Brain Tumor Segmentation Using Cascaded Anisotropic Convolutional Neural Networks. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*; Springer: Cham, Switzerland, 2017; pp. 178–190.

20. Myronenko, A. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*; Springer: Cham, Switzerland, 2018; pp. 311–320.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *31st International Conference on Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2017.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations, Virtual Event*, 3–7 May 2021.
23. 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 *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*; Springer: Cham, Switzerland, 2022; Volume 12962, pp. 272–284. [https://doi.org/10.1007/978-3-031-08999-2\\_22](https://doi.org/10.1007/978-3-031-08999-2_22).
24. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning, Virtual*, 18–24 July 2021; pp. 8748–8763.
25. 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* **2020**, *36*, 1234–1240. <https://doi.org/10.1093/bioinformatics/btz682>.
26. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc.* **2021**, *3*, 2. <https://doi.org/10.1145/3458754>.
27. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.H.; Jin, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In *Proceedings of the Clinical Natural Language Processing Workshop*; Association for Computational Linguistics: Stroudsburg PA USA, 2019; pp. 72–78.
28. Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C.D.; Langlotz, C.P. Contrastive Learning of Medical Visual Representations from Paired Images and Text. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, Durham, NC, USA, 5–6 August 2022; pp. 2–25.
29. 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 *IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 3942–3951.
30. Wang, Z.; Wu, Z.; Agarwal, D.; Sun, J. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg PA USA, 2022; pp. 3876–3887.
31. Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Scientific Image-Text Pairs. *arXiv* **2023**, arXiv:2303.00915.
32. Karimi, D.; Salcudean, S.E. Reducing The Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks. *IEEE Trans. Med. Imaging* **2020**, *39*, 499–513. <https://doi.org/10.1109/TMI.2019.2930068>.
33. Oppenheim, A.V.; Schaffer, R.W.; Buck, J.R. *Discrete-Time Signal Processing*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 1999.
34. Bracewell, R.N. *The Fourier Transform and Its Applications*, 3rd ed.; McGraw-Hill: New York, NY, USA, 1999.
35. Wang, H.; Wu, X.; Huang, Z.; Xing, E.P. High-Frequency Component Helps Explain The Generalization of Convolutional Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2020; pp. 8684–8694.
36. Wang, K.; Li, Z.; Wang, H.; Liu, S.; Pan, M.; Wang, M.; Wang, S.; Song, Z. Improving Brain Tumor Segmentation with Anatomical Prior-Informed Pre-Training. *Front. Med.* **2023**, *10*, 1211800. <https://doi.org/10.3389/fmed.2023.1211800>.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.