Simultaneous Tumor Delineation and Report Generation from Brain MR Images

A dissertation submitted in partial fulfillment of the requirements for the degree of

Master of Technology

in

Computer Science

by

Adish Mallik

Roll no. - **CS2403**

under the supervision of

Prof. Pradipta Maji

Machine Intelligence Unit (MIU)

Indian Statistical Institute, Kolkata

June, 2026

Certificate

I certify that the project work titled "**Simultaneous Tumor Delineation and Report Generation from Brain MR Images**" authored by Adish Mallik, has been conducted under my supervision and guidance. The thesis represents the original work of the candidate, demonstrating thorough research and investigation. The quality of the thesis meets the expectations for the Master of Technology program in Computer Science, and I highly recommend its submission for evaluation.

Pji 10.06.2026

Prof. Pradipta Maji
Machine Intelligence Unit (MIU)
Indian Statistical Institute
Kolkata - 700108, India

Acknowledgement

I would like to express my sincere gratitude to Prof. Pradipta Maji, my supervisor at the Machine Intelligence Unit (MIU) of the Indian Statistical Institute, Kolkata, for his guidance, support, and valuable feedback throughout this project. I am also thankful to all faculty members, colleagues, and friends who supported me during the course of this work.

Abstract

Brain tumor analysis is an important application of medical image computing, where accurate segmentation and interpretation of tumor regions can support diagnosis and treatment planning. However, existing methods often address tumor segmentation and radiology report generation as separate tasks. Moreover, one of the major challenges in report generation tasks from MRI is accurate tumor localization. Most models fail to locate the lobe and hemisphere in which the tumor is located, causing incorrect report generation. In this regard, a unified 3D vision-language framework is proposed for simultaneous brain tumor segmentation and report generation from multi-modal MRI. Given T1, T2, T1C, and FLAIR volumes, the proposed model predicts clinically meaningful tumor regions and generates a textual description of tumor location and appearance. In order to encourage consistency between the predicted segmentation and generated report, the proposed model judiciously integrates a Swin UNETR-based 3D encoder-decoder, a multi-scale lesion tokenizer, auxiliary clinical grounding heads, and an iterative cross-modal refinement module. Moreover, anatomy and laterality heads are introduced, which provide clinical hints to the LLM, allowing better tumor localization. Further, an iterative refinement is incorporated so that the generated report and segmented outputs can refine each other, finally producing better outputs. Extensive experimentation on BraTS2020 and TextBraTS data sets shows that the proposed model achieves a mean Dice of 81.60% and HD95 of 6.21. In addition, the model achieves a BERTScore-F1 of 0.9226, clinical laterality F1 of 0.8459, clinical anatomy F1 of 0.7539 and clinical pathology F1 of 0.9976. These results indicate that the proposed framework can generate clinically meaningful reports while accurately localizing tumor regions and maintaining strong alignment between segmentation and textual interpretation.

Keywords: Brain MRI, tumor segmentation, radiology report generation, vision-language model, 3D Deep learning, cross-modal refinement

Contents

List of Figures

Chapter 1

Introduction

Brain tumor analysis is currently one of the most important applications in medical imaging because accurate localization, characterization, and interpretation help in diagnosis and treatment planning. It provides a preliminary idea of whether surgery is required or whether radiotherapy may be sufficient. Magnetic resonance imaging (MRI) is the standard imaging modality for brain tumor assessment. It provides rich anatomical and pathological information through multiple sequences such as T1, T1C, T2, and FLAIR. The core task in tumor analysis is brain tumor segmentation, where the objective is to automatically delineate clinically meaningful tumor sub-regions such as Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). These sub-regions provide information about edema, necrosis, and enhancement, which are essential for clinical evaluation.

However, segmentation alone does not fully represent the way radiologists interpret MRI scans. In real clinical practice, radiologists not only outline tumor boundaries but also generate textual reports describing lesion location, signal characteristics, laterality, edema, necrosis, mass effect, ventricular compression, and other clinically meaningful observations. Therefore, a more complete artificial intelligence system for brain tumor analysis should be capable of both producing accurate segmentation masks and generating structured reports.

Most common brain tumor segmentation models are designed as image-only systems. They take 2D or 3D images as input and produce segmentation masks as output. U-Net [2], nnU-Net [3], and Swin UNETR [1] are examples of image-only systems that show strong performance on benchmark datasets. However, recent research on the incorporation of radiological reports has shown that text can improve the quality of segmentation. One recent model, TextBraTS [4], demonstrates that radiological text can provide useful semantic guidance for improving segmentation performance.

On the other hand, report generation from 2D and 3D medical images has become an important research direction due to the rapid progress of vision-language models. Med3DVLM [5] and MedGemma [6] are some recent works in this area. Some methods process MRI scans slice by slice, while others take the entire 3D MRI scan as input to generate the radiology report.

Despite these advances, segmentation and report generation are often studied as separate tasks. Text-guided segmentation methods usually use reports as input to improve mask prediction, while report generation models usually generate text from images without explicitly producing segmentation masks.

The objective of this project is to develop a unified model for brain MRI analysis that jointly performs tumor segmentation and radiology report generation. Given multi-modal 3D MRI input, the model aims to predict clinically relevant tumor sub-regions such as WT, TC, and ET, while also generating a textual report describing the tumor location and appearance. Furthermore, we use a refinement technique that allows the segmentation task and the report generation task to benefit from each other, thereby improving the final outputs.

Our Contribution :

This work makes three main contributions:

1. We propose a unified 3D vision-language framework that is capable of both segmenting tumor regions and generating reports simultaneously from multi-modal MRI volumes.

2. We introduce anatomy and laterality heads to improve tumor localization. These heads, along with geometry extraction and anatomical query generation, provides necessary information to the language model, helping it generate reports that more accurately describe the tumor hemisphere and anatomical lobe.

3. We design an iterative cross-modal refinement module that allows the segmentation and report-generation branches to exchange information. This helps the two branches to learn from each other, thereby improving their outputs.

Chapter 2

Background and Related Works

2.1 Background

Magnetic Resonance Imaging (MRI) is a medical imaging technique used to produce images of anatomical structures. It is widely used for brain tumor analysis because it provides rich soft-tissue contrast and captures different structural and pathological properties of the brain. In the proposed work, we use four MRI modalities: T1, T1C, T2, and FLAIR.

T1-weighted MRI provides detailed anatomical information about brain structures. It can help in locating major brain structures such as gray matter, white matter, ventricles, and cerebrospinal fluid. However, edema and tumor regions are often poorly visible on T1-weighted images. This modality is often used for locating tumor core.

T1C, or contrast-enhanced T1-weighted MRI, is acquired after injecting contrast agent before imaging. It is useful for showing regions where the blood-brain barrier is disrupted and is helpful for segmenting Tumor Core and Enhancing Tumor. However, edema is poorly visible in this modality.

In **T2-weighted MRI**, fluid-containing regions appear bright. Thus it is useful for detecting abnormalities that contain increased water content like edema and segmenting the whole tumor extent. However, enhancing tumor regions and tumor core are poorly visible in this modality.

FLAIR, or Fluid-Attenuated Inversion Recovery, suppresses the signal from cerebrospinal fluid while preserving abnormal hyperintense regions. FLAIR is particularly useful for segmenting edema, non-enhancing tumor components, and the overall tumor extent, while it is less effective for accurately segmenting enhancing tumor components.

Each modality provides complementary information. T1 and T1C are useful for anatomical structure and enhancement, while T2 and FLAIR are useful for edema and tumor spread. Therefore, combining all four modalities allows one to learn a more complete representation of tumor appearance, location, and extent.

2.2 Related Works

Existing studies are reviewed along two main directions: segmentation mask generation and report generation.

2.2.1 Segmentation Mask Generation

Medical image segmentation is one of the most established tasks in computer-aided diagnosis. In brain tumor analysis, the BraTS benchmark has played an important role in standardizing this task by defining clinically meaningful tumor sub-regions, commonly including Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). Segmenting these regions is challenging because tumor boundaries are often irregular, heterogeneous, and visually ambiguous across different MRI modalities. As a result, brain tumor segmentation has gradually evolved from classical image-processing techniques to deep learning-based architectures that can learn hierarchical spatial features from multi-modal MRI volumes.

Early deep learning methods for medical image segmentation were largely based on convolutional neural networks, especially U-Net-based architectures. U-Net [2] introduced an encoder-decoder structure with skip connections, which allowed the model to combine low-level spatial details with high-level semantic information. However, since the original U-Net [2] operates on 2D slices, it cannot fully capture the 3D spatial context present in volumetric medical images.

To address this limitation, later U-Net variants extended the architecture to process volumetric data directly. For example, 3D U-Net [7] replaced 2D operations with 3D convolutions, 3D max-pooling, and 3D up-convolutions. This enabled the network to process full 3D image volumes instead of independent 2D slices, allowing it to exploit inter-slice spatial continuity.

Another important development in medical image segmentation was nnU-Net [3], which proposed a self-configuring framework rather than a completely new network architecture. nnU-Net [3] automatically adapts the segmentation pipeline according to the properties of a given dataset, including pre-processing, network configuration, training strategy,

inference, and post-processing. However, like most classical segmentation models, nnU-Net [3] is primarily image-only. It uses MRI volumes as input and predicts segmentation masks without directly incorporating textual clinical descriptions or radiological prior knowledge.

To overcome the limited ability of CNNs to capture long-range dependencies, transformer-based models were later introduced for brain tumor segmentation. One important example is Swin UNETR [1], which uses a Swin Transformer [8] as the encoder and a CNN-based decoder for mask reconstruction. The input MRI volume is divided into patches and represented as a sequence of embedding, allowing the model to learn both local and global contextual information. Swin UNETR [1] uses a hierarchical transformer encoder to extract features at multiple resolutions using shifted-window self-attention.

Following this direction, several 2D medical segmentation models have been proposed that leverage medical reports or textual prompts to improve segmentation performance. For example, LViT [9] incorporates textual annotations with 2D image features to guide mask prediction. MedCLIP-SAM [10] combines medical vision-language representation learning with SAM-style segmentation to support text-guided mask generation. Similarly, MedCLIPSeg [11] adopts a CLIP-based segmentation framework, where image patches are aligned with textual prompts to produce dense segmentation masks.

More recently, TextBraTS [4] introduced a text-guided approach for volumetric brain tumor segmentation. It proposed a dataset consisting of 3D MRIs and their corresponding reports, and developed a model that used reports to improve segmentation mask prediction. The model uses Swin UNETR [1] as the image backbone and BioBERT [12] as the text encoder. The extracted text features are mapped into the image feature space and fused with image features using a bidirectional cross-attention fusion module.

2.2.2 Report Generation

Another important task in the medical domain is the automated generation of clinical reports. The goal is to generate factual clinical reports from a given medical image. These reports should correctly describe the location and appearance of tumor regions while also being grammatically correct and easy to understand. A good model should avoid hallucinating findings that are not present in the image and should correctly report the actual structures and abnormalities present in the brain.

Early models such as ConVIRT [13] use an image encoder to extract visual features and a text encoder to extract textual features. They use contrastive learning to learn visual representations from paired image-text data. Following this, MedCLIP [14] used

a similar CLIP-like framework and replaced the InfoNCE loss with a semantic matching loss, enabling training on unpaired image and text data.

Report generation models generally follow an encoder-decoder structure. The encoder is commonly composed of CNNs or Transformers to extract image features. These features are then passed to a language-generating module, such as an LSTM or, more recently, a large language model. Examples include R2GenGPT [15], Med-Flamingo [16], LLaVA-Med [17], CT2Rep [18], and CT-CHAT [19].

MedGemma [6] is a family of medical vision-language foundation models based on Gemma 3, available in 4B and 27B variants. It is designed for medical image and text understanding, mainly for 2D medical images.

Med3DVLM [5] addresses the difficulty of applying vision-language models to volumetric medical images, where 3D scans are computationally expensive and image-text alignment is more challenging than in 2D images. The model consists of a 3D vision encoder, a multimodal projector, and a large language model.

Another recent work in report generation is Brain3D [20], which focuses specifically on automated radiology report generation from 3D brain tumor MRI. Brain3D converts a pretrained 2D medical vision encoder into a native 3D encoder through weight inflation. The extracted 3D visual tokens are compressed and projected into the embedding space of a causal language model.

Chapter 3

Proposed Method

3.1 Overview

We present a unified 3D vision-language framework for simultaneous brain tumor segmentation and radiology report generation from multi-modal MRI. Given the four MRI modalities, T1, T2, T1C, and FLAIR, we first use a Swin UNETR encoder-decoder to predict region-wise tumor masks for the tumor core, whole tumor, and enhancing tumor. The Swin UNETR [1] encoder is further used to extract hierarchical 3D visual features. A multi-scale lesion tokenizer then converts these image features into visual tokens, which are provided to the language model as input along with the textual prompt.

To improve clinical grounding, we further introduce laterality and anatomy heads. The outputs of these two heads provide information about the tumor location and are added to the language model prompt in the form of clinical hints. In addition, we use an iterative cross-modal refinement module that feeds language-derived features back into the image representation. This design allows the segmentation and report generation branches to interact with each other, encouraging better consistency between the predicted tumor mask and the generated clinical description.

3.2 Problem Formulation

Let

$$X \in \mathbb{R}^{4\times D\times H\times W} \tag{3.1}$$

be a 3D brain MRI volume, where the four input channels correspond to T1, T2, T1C, and FLAIR modalities. The objective is to jointly perform brain tumor segmentation

and radiology report generation from this input volume.

For the segmentation task, the model predicts a region-wise tumor mask

$$\hat{S} \in \mathbb{R}^{3\times D\times H\times W}, \tag{3.2}$$

where the three output channels correspond to the three tumor sub regions: Tumor Core (TC), Whole Tumor (WT), and Enhancing Tumor (ET).

For the report generation task, the model generates a textual radiology report

$$\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T\}, \tag{3.3}$$

where $\hat{y}_t$ represents the token generated at time step t. The generated report is conditioned on lesion-aware visual tokens extracted from the input MRI volume and the predicted tumor regions. Therefore, the overall objective is to learn a joint mapping

$$f_\theta : X \to (\hat{S}, \hat{Y}), \tag{3.4}$$

where θ denotes the trainable parameters of the complete vision-language framework.

During training, the model is supervised using the ground-truth tumor segmentation mask S, the reference report Y, and auxiliary clinical labels for laterality and anatomical location. We use the ground truth segmentation mask during training and the predicted segmentation mask during inference to create the visual tokens.

3.3 Segmentation Module

We use the Swin UNETR [1] encoder as the 3D visual feature extractor. The encoder produces a bottleneck feature and several intermediate feature maps, which are passed to the Swin UNETR [1] decoder for segmentation. During the refinement stage, the bottleneck representation is updated using text-derived features extracted from the report decoder.

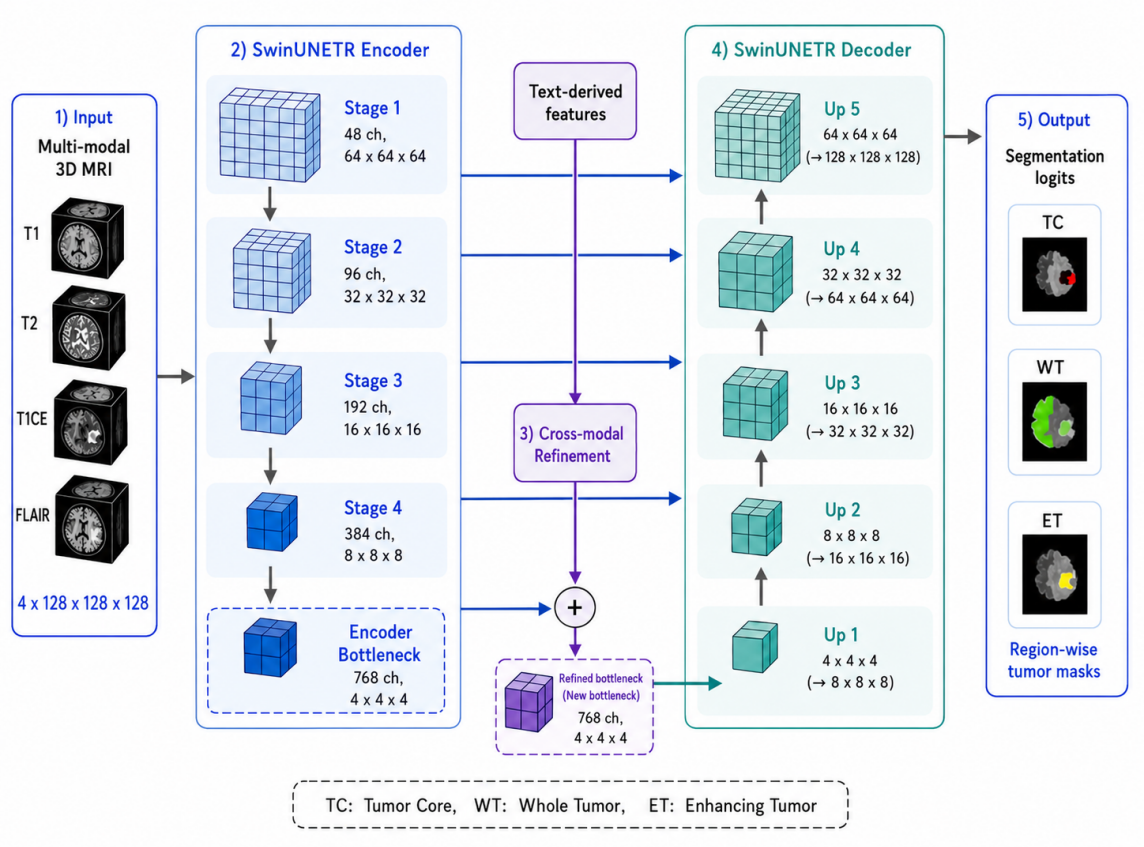

Figure 3.1: Segmentation module. The Swin UNETR [1] encoder extracts hierarchical visual features from the multi-modal MRI input. The decoder predicts region-wise tumor masks for TC, WT, and ET.

3.4 Multi-Scale Lesion Tokenizer

The model uses feature maps from three different levels of the Swin UNETR [1] backbone. Fine tokens are created from Stage 2 encoder features and capture local details such as tumor boundaries and small structures. Mid-level tokens are created from Stage 4 encoder features and capture regional tumor context. Global tokens are created from the bottleneck features and represent the overall image-level context.

Let $\mathbf{F}^{(s)} \in \mathbb{R}^{C_s \times D_s \times H_s \times W_s}$ denote the encoder feature map at scale s, where C_s is the number of channels and (D_s, H_s, W_s) is the spatial resolution of the feature map. Let $\mathbf{L} \in \mathbb{R}^{3 \times D \times H \times W}$ denote the segmentation logit map predicted by the segmentation decoder, where the three channels correspond to tumor core (TC), whole tumor (WT), and enhancing tumor (ET).

First, the feature map at each scale is projected to a common embedding dimension using a $1 \times 1 \times 1$ convolution. The projected feature is then normalized and passed through a

non-linear GELU activation function:

$$\mathbf{Z}^{(s)} = \phi\left(\mathrm{Norm}\left(\mathrm{Conv}^{(s)}_{1\times1\times1}\left(\mathbf{F}^{(s)}\right)\right)\right), \tag{3.5}$$

where $\mathrm{Norm}(\cdot)$ denotes the normalization layer and $\phi(\cdot)$ denotes the activation function.

The segmentation logits are then converted into soft region probabilities using the sigmoid function:

$$\mathbf{P} = \sigma(\mathbf{L}), \tag{3.6}$$

where $\mathbf{P} \in \mathbb{R}^{3\times D\times H\times W}$ represents the predicted soft tumor-region probability map.

Since the spatial resolution of $\mathbf{P}$ may be different from the resolution of $\mathbf{Z}^{(s)}$, the probability map is resized using adaptive average pooling:

$$\mathbf{G}^{(s)} = \mathrm{AAP}\left(\mathbf{P}, (D_s, H_s, W_s)\right), \tag{3.7}$$

where $\mathrm{AAP}(\cdot)$ denotes adaptive average pooling. The resulting $\mathbf{G}^{(s)}$ is used as the soft lesion-guidance map for scale s.

The projected feature maps from Stage 2 and Stage 4 are then modulated using the lesion-guidance map:

$$\widetilde{\mathbf{Z}}^{(s)} = \mathbf{Z}^{(s)} \odot \left(1 + \alpha\mathbf{G}^{(s)}\right), \tag{3.8}$$

where $\odot$ denotes element-wise multiplication and α is a scaling coefficient that controls the strength of lesion guidance. In the proposed implementation, α is set to 0.5. For the bottleneck stage, lesion guidance is not used, and only $\mathbf{Z}^{(s)}$ is used.

The final output is then passed through another adaptive average pooling operation that reduces the spatial size to $3\times3\times3$. The pooled output is flattened to obtain 27 tokens for each scale. We also add learnable positional encoding and scale encoding to each token.

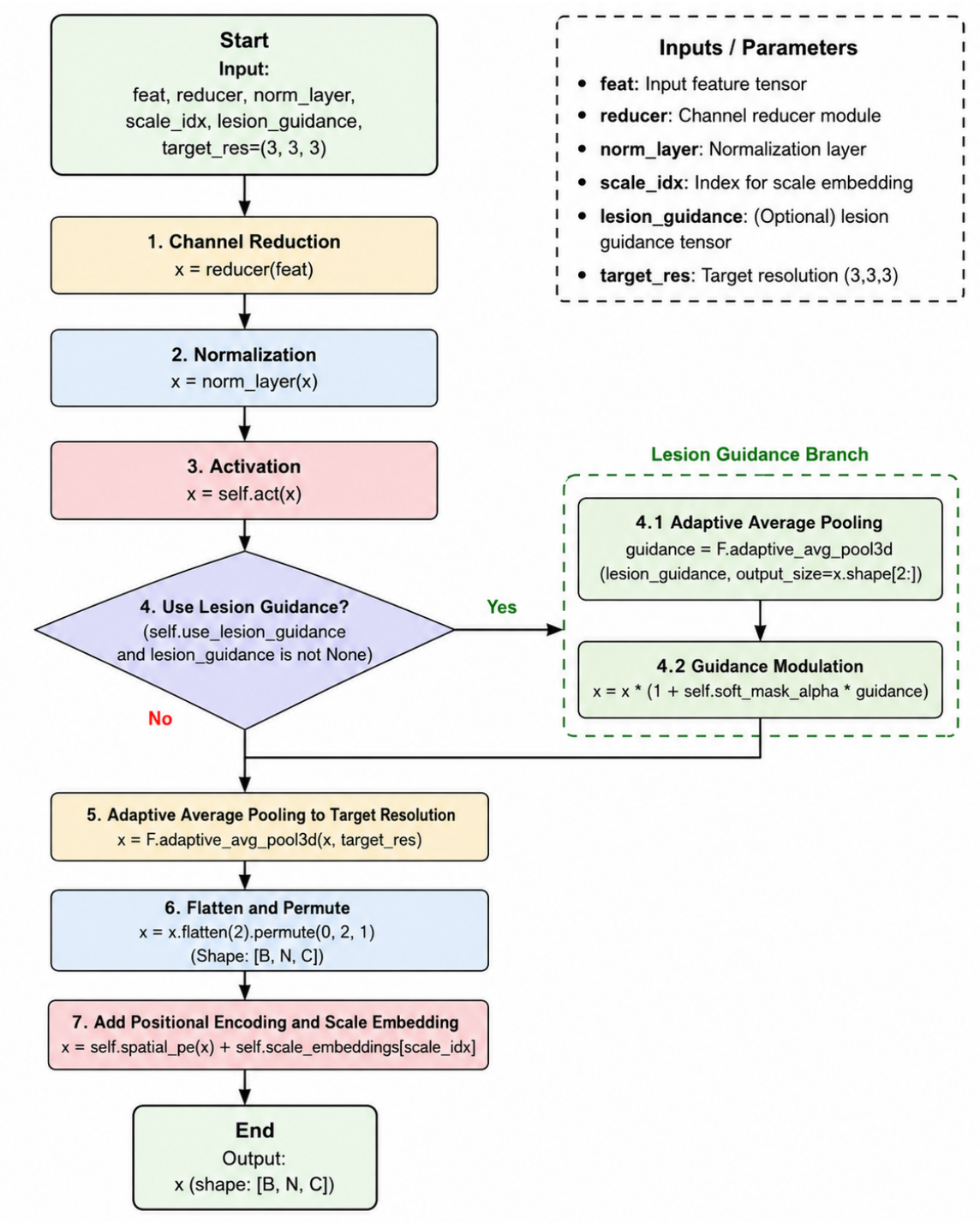

Figure 3.2: Visual Memory Token Creation using encoder features from multiple levels

The model finally outputs 27 fine tokens, 27 mid-level tokens, and 27 global tokens, which are concatenated to form 81 visual memory tokens.

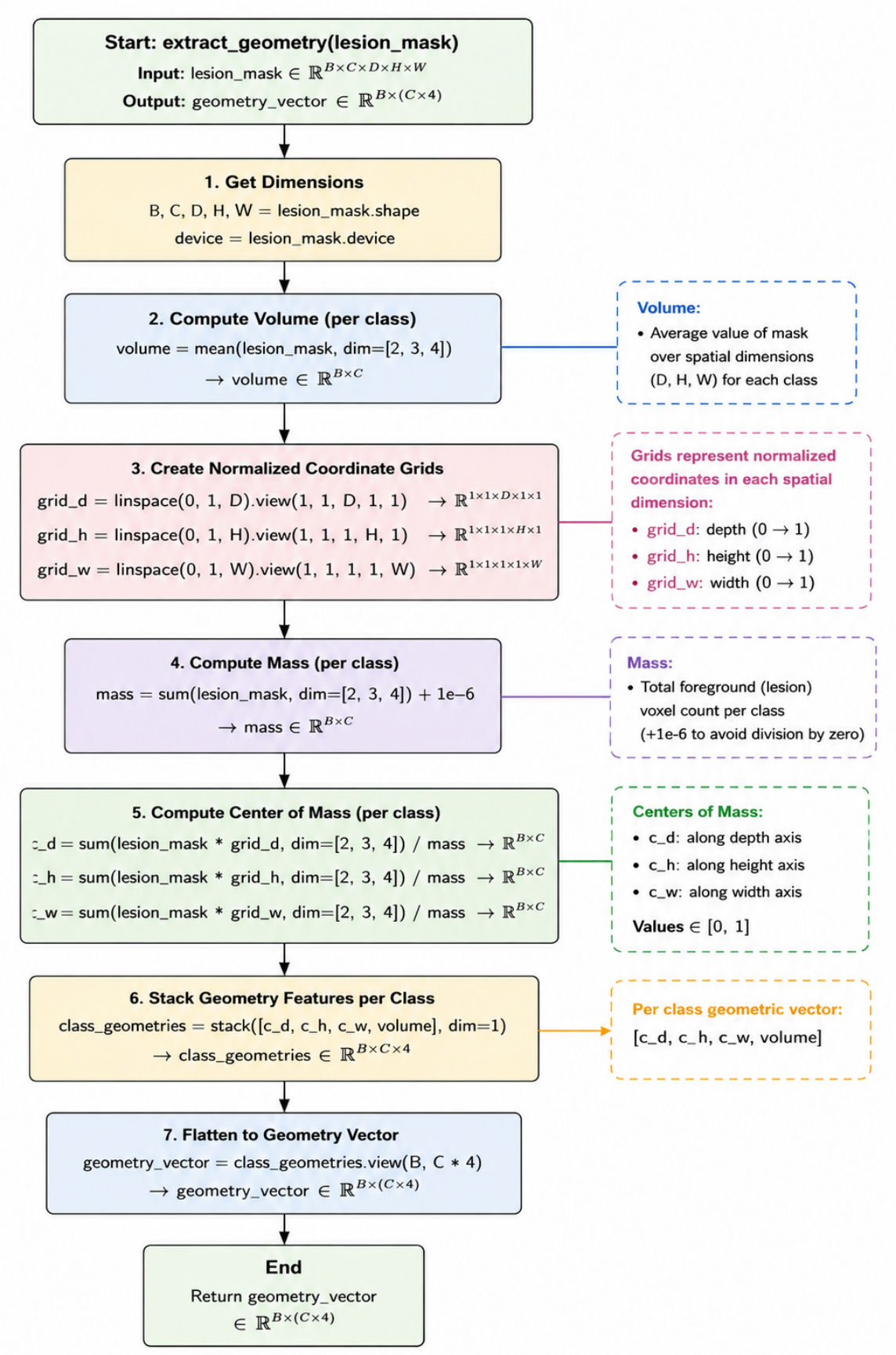

Figure 3.3: Geometry Extraction Module for extracting center of mass and volume for each anatomical region

Further, to provide explicit spatial information, we extract geometry from the predicted lesion masks. For each tumor region (edema, necrosis and enhancing tumor), the model calculates the center of mass along the depth, height, and width directions, along with the normalized tumor volume. Therefore, each tumor class is represented using four geometry values: location in 3D space and size.

Let $\mathbf{M} \in \mathbb{R}^{B\times C\times D\times H\times W}$ denote the segmentation mask, where B is the batch size, C is the number of segmentation mask channels, and (D, H, W) represents the spatial dimensions of the 3D volume. From the segmentation module, we get 3 channel segmentation mask, corresponding to Whole Tumor (WT), Tumor Core (TC) and Enhancing Tumor (ET) as output. We use this 3 segmentation masks to compute masks for edema (WT - TC) and necrosis (TC - ET). We extract geometry for these three anatomical features (edema, necrosis and enhancing tumor).

To compute the spatial location of each anatomical region, we define normalized coordinate grids along the depth, height, and width axes:

$$r_d = \frac{d}{D-1}, \qquad r_h = \frac{h}{H-1}, \qquad r_w = \frac{w}{W-1}, \tag{3.9}$$

where r_d, r_h, and r_w lie in the range $[0, 1]$. These normalized coordinates make the geometric representation independent of the absolute image size.

The total mass for class c is then computed as:

$$m_c = \sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{w=1}^{W} M_{c,d,h,w} + \epsilon, \tag{3.10}$$

where ϵ is a small constant added to avoid division by zero when the predicted segmentation region is empty.

The center of mass for the class c is then computed as the mask-weighted average of the normalized voxel coordinates. The depth coordinate of the center of mass is given by:

$$\bar{d}_c = \frac{\sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{w=1}^{W} M_{c,d,h,w} * r_d}{m_c}. \tag{3.11}$$

Similarly, we also calculated the height and width coordinate if the center of mass for each class. The final center of mass for class c is then represented as: $\left[\bar{d}_c, \bar{h}_c, \bar{w}_c, volume\right]$, where *volume* is computed by averaging the segmentation mask over all spatial locations for each channel.

Once we have the visual memory tokens and geometry tokens, we create the queries for the language model.

Let, $\mathbf{V} \in \mathbb{R}^{B \times N_{\mathrm{m}} \times C}$ denote the visual memory, where B is the batch size, N_{m} is the number of visual memory tokens (in the proposed case, it is 81), and C is the embedding dimension (in the proposed case, it is 512). The visual memory is obtained by concatenating the fine, mid-level, and global tokens.

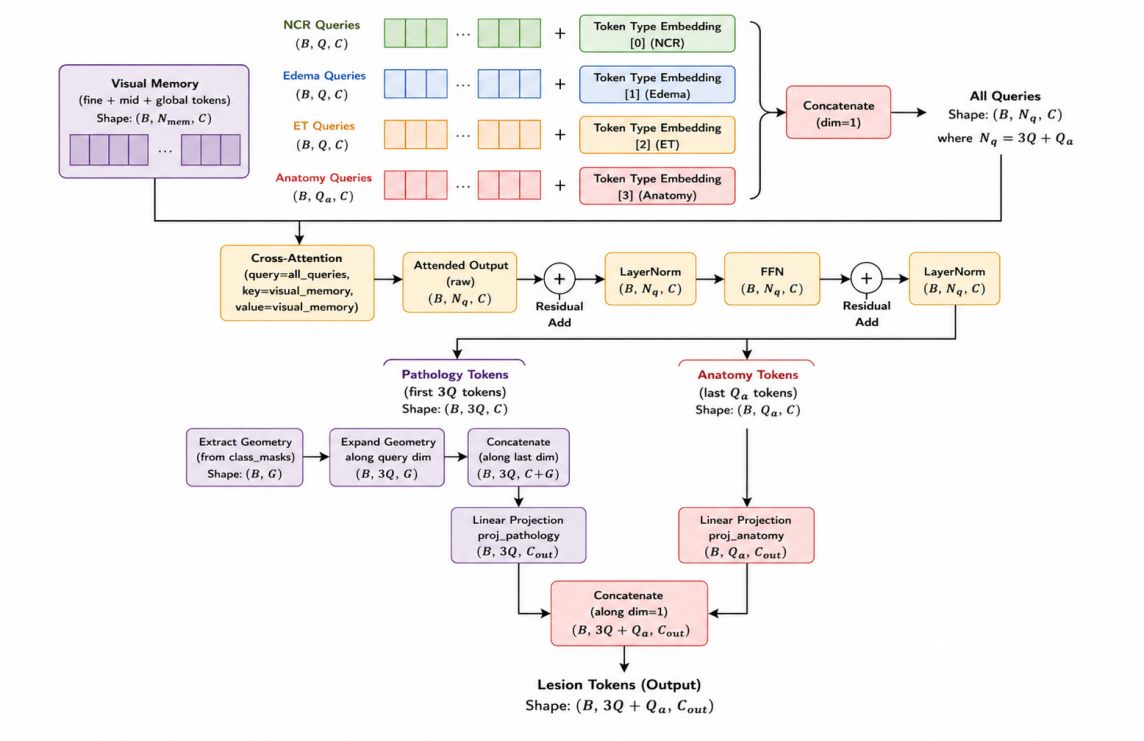

Figure 3.4: Query Creation Module. The final lesion tokens output is prepended to the prompt before sending it to the language model.

To extract both lesion-specific and full anatomical information from this visual memory, we define separate learnable query groups for different tumor-related concepts. Let, $\mathbf{Q}_{\mathrm{NCR}} \in \mathbb{R}^{B \times Q \times C}$, $\mathbf{Q}_{\mathrm{ED}} \in \mathbb{R}^{B \times Q \times C}$, $\mathbf{Q}_{\mathrm{ET}} \in \mathbb{R}^{B \times Q \times C}$ and $\mathbf{Q}_{\mathrm{A}} \in \mathbb{R}^{B \times Q_a \times C}$ denote the learnable query tokens for necrosis, edema, enhancing tumor and full anatomy respectively. Q is the number of queries for tumor type and Q_a is the number of queries for anatomical information. In the proposed case, $Q = Q_a = 32$. Further, to differentiate between the query types, a learnable token type embedding is added to each query group. We then concatenate the four query groups along the token dimension. This concatenated query tokens attend to the visual memory through a cross-attention module. This cross-attention allows each learnable query to retrieve relevant information from the visual memory.

We then apply a residual connection and layer normalization and pass the output through

a feed-forward network. The output of the feed-forward network is again added to the residuals and passed through a layer normalization.

$$\mathbf{T} = \text{LayerNorm}\left(\mathbf{H}_1 + \text{FFN}\left(\mathbf{H}_1\right)\right), \qquad \mathbf{H}_1 = \text{LayerNorm}\left(\mathbf{Q}_{\text{all}} + \mathbf{A}\right). \tag{3.12}$$

where $\mathbf{Q}_{\text{all}}$ denotes the concatenated learnable query tokens for necrosis, edema, enhancing tumor, and anatomy, and $\mathbf{A}$ denotes the cross-attention output obtained by attending $\mathbf{Q}_{\text{all}}$ to the visual memory tokens.

The refined token sequence T is then split into pathology tokens and anatomy tokens. The first 3Q tokens correspond to pathology-related tumor tokens, while the remaining Q_a tokens correspond to anatomy tokens. Further, we concatenate the geometry tokens with the pathology tokens and project the concatenated tokens to LLM dimension (in the proposed case, it is 2048). A separate projection layer projects the anatomy tokens to the LLM dimension. The final lesion tokens are formed by concatenating the projected pathology tokens and projected anatomy tokens. These tokens contain pathology-specific, anatomy-aware, and explicit geometric information about lesion location and extent.

3.5 Clinical Grounding Heads

We use two neural network models to predict the laterality and the anatomy present in the MRI. The laterality head predicts whether the tumor is located on the left, right, or bilateral side of the brain. The anatomy head predicts the anatomical regions involved by the tumor. It is a multi-label classifier because a tumor can involve more than one region. The predicted anatomy labels include frontal, parietal, temporal, occipital, cerebellum, and ventricle/periventricular regions.

Let $\mathbf{F}_b \in \mathbb{R}^{B\times C\text{in}\times D\times H\times W}$ denote the bottleneck feature extracted by the Swin UNETR encoder, where B is the batch size, C_{in} is the number of input channels of the head, and (D, H, W) is the spatial resolution of the bottleneck feature. In the proposed case, $C_{\text{in}} = 768$

First, global spatial information is aggregated from the bottleneck feature $\mathbf{F}_b$ using adaptive average pooling and then flattened:

$$\mathbf{x} = \text{Flatten}\left(\text{AAP}_{3D}\left(\mathbf{F}_b\right)\right), \qquad \mathbf{x} \in \mathbb{R}^{B\times C_{\text{in}}}. \tag{3.13}$$

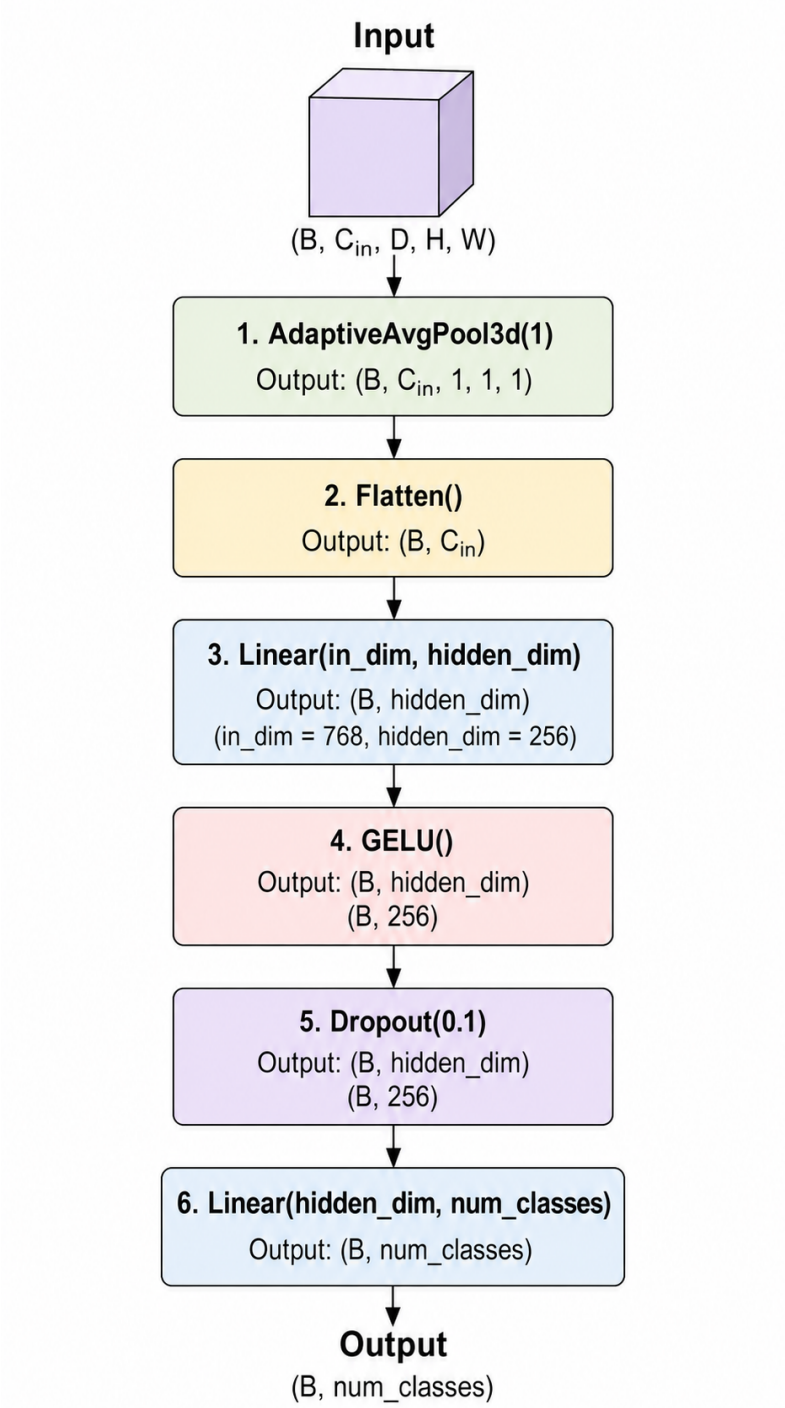

Figure 3.5: Clinical heads module. The bottleneck feature is passed through laterality and anatomy heads to generate clinical hints, which are added to the language model prompt. We use two separate modules that share the same structure except for the number of classes for predicting anatomy and laterality.

The flattened feature $\mathbf{x}$ is passed through a fully connected linear layer with GELU activation and dropout of 0.1 to produce the output logits:

$$\mathbf{o} = \text{Dropout}\left(\text{GELU}\left(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1\right)\right) \qquad \mathbf{o} \in \mathbb{R}^{B \times C_{\text{out}}}. \tag{3.14}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C\text{in} \times C_h}$, and $\mathbf{b}_1 \in \mathbb{R}^{C_h}$ are learnable parameters. Here, C_h is the hidden dimension, and in the proposed implementation $C_h = 256$.

Finally, the output logits are computed using another fully connected layer:

$$\mathbf{o} = \mathbf{h}\mathbf{W}_2 + \mathbf{b}_2, \tag{3.15}$$

where $\mathbf{W}_2 \in \mathbb{R}^{C_h \times C_{\text{out}}}$ and $\mathbf{b}_2 \in \mathbb{R}^{C_{\text{out}}}$ are the parameters of the second linear layer, and C_{out} is determined by the number of classes to be predicted.

We used the same neural network structure but separate networks for both the anatomy and laterality heads. The number of classes predicted by laterality head is 3 corresponding to left, right and bilateral. The number of classes predicted by anatomy head is 6 corresponding to frontal, parietal, temporal, occipital, cerebellum, and ventricle/periventricular regions. The predicted probability for each anatomical region is computed using the sigmoid function and are used to create clinical grounding hints, such as: "Laterality: left. Anatomy: frontal, temporal." These hints are added to the prompt given to the report decoder.

The ground truth for the two heads are created by extracting the relevant contents from the ground truth reports.

3.6 Vision-Conditioned Report Generation

Using the outputs of laterality heads we first create the prompt.The prompt contains a system message and user message. The outputs from the laterality and anatomy heads is given as hints in the prompt. The clinical hint follows the following format: Laterality: laterality. Anatomy: anatomy list. For example: Laterality: left. Anatomy: frontal, temporal. This means that the tumor is located on the left side of the brain and involves the frontal and temporal regions

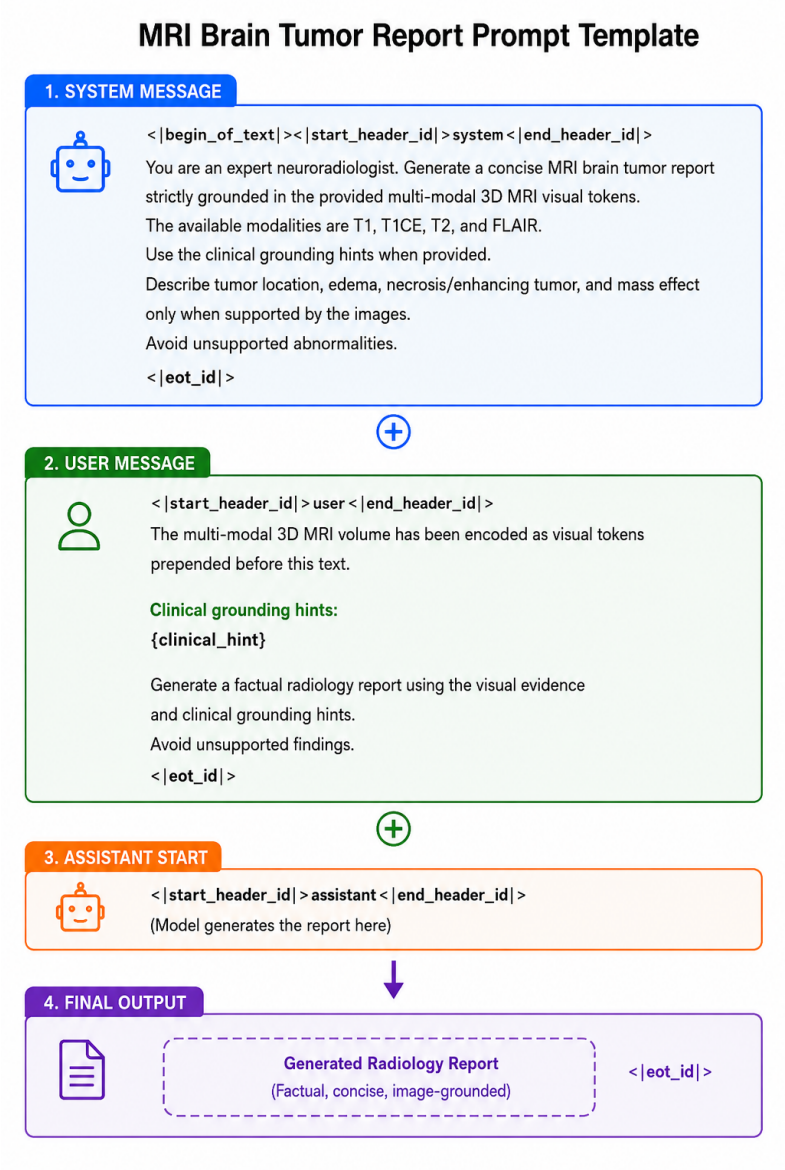

Figure 3.6: Prompt construction for report generation. The lesion tokens produced by the tokenizer are prepended before the textual prompt and passed to the language model. The prompt contains a system instruction, a user message, and clinical grounding hints obtained from the laterality and anatomy heads. The language model then generates the final radiology report conditioned on both the visual lesion tokens and the textual prompt.

The lesions tokens are then prepended to the prompt and is given input to the LLM.

3.7 Cross-Modal Iterative Refinement

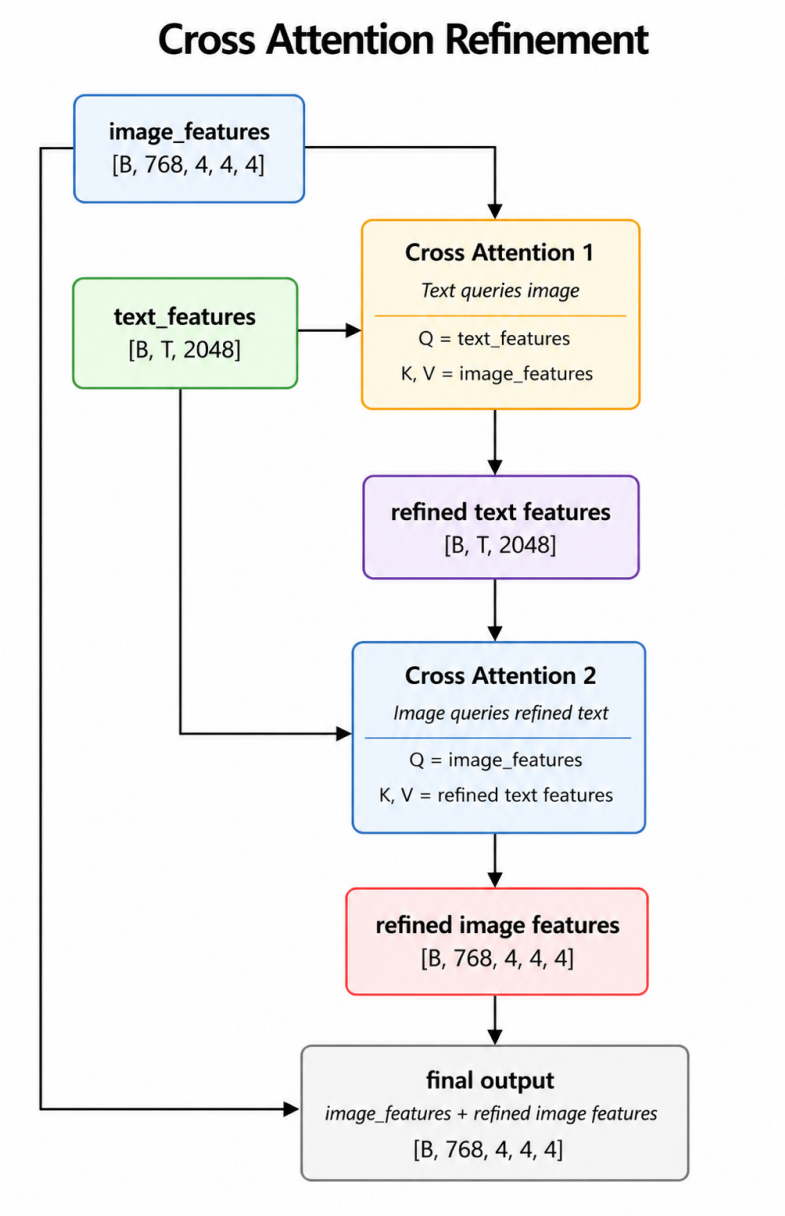

Figure 3.7: Cross-modal refinement module. Text-derived features first attend to the image bottleneck features, and the refined text features are then used to update the image representation. The final output is obtained by adding the refined image features to the original image bottleneck.

This stage links segmentation and report generation via iterative feedback. Initially, the decoder uses MRI and bottleneck features to predict segmentation. This output is then used to create the prompt for report generation. A cross-attention refinement

module then uses these report features to refine the image bottleneck feature, allowing text information to enhance the image representation. The refined bottleneck is then used by the segmentation decoder in the next iteration, improving segmentation with report-level information in each step. Finally, the prompt tokens are passed to the report decoder with ground-truth labels to calculate the report-generation loss.

The cross-attention refinement module we use is similar to TextBraTS [4]. The text features first attend to the image bottleneck features to obtain image-aware text representations, and the image features then attend back to these refined text features. The resulting refined image representation is added to the original bottleneck through a residual connection.

3.8 Training Objective

3.8.1 Loss Functions

The overall training objective combines segmentation, report generation, and auxiliary clinical grounding losses:

$$\mathcal{L} = \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{LLM}}\mathcal{L}_{\text{LLM}} + \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_{\text{anat}}\mathcal{L}_{\text{anat}}, \tag{3.16}$$

where $\mathcal{L}_{\text{seg}}$ is the segmentation loss, $\mathcal{L}_{\text{LLM}}$ is the language-modeling loss, $\mathcal{L}_{\text{lat}}$ is the laterality classification loss, and $\mathcal{L}_{\text{anat}}$ is the anatomy classification loss.

For segmentation, we use a combination of Dice loss and Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}}. \tag{3.17}$$

Dice loss encourages overlap between the predicted tumor regions and the ground-truth masks, while BCE provides voxel-wise supervision by penalizing incorrect predictions at each voxel.

For laterality prediction, which is a three-class classification task, we use the standard cross-entropy loss:

$$\mathcal{L}_{\text{lat}} = \mathcal{L}_{\text{CE}}. \tag{3.18}$$

For anatomy prediction, which is a multi-label classification task, we use Binary Cross-Entropy loss:

$$\mathcal{L}_{\text{anat}} = \mathcal{L}_{\text{BCE}}. \tag{3.19}$$

The laterality and anatomy losses are weighted by coefficients λ_{lat} and λ_{anat}, respectively.

For report generation, we use the standard causal language-modeling loss. Given a report token sequence $\{y_1, y_2, \ldots, y_N\}$, the model is trained to predict the next token at each position:

$$\mathcal{L}_{\text{LLM}} = -\sum_{t=1}^{N} \log P(y_t \mid y_{<t}). \tag{3.20}$$

The lesion tokens and prompt tokens do not correspond to target report words. Therefore, their labels are set to -100, causing the cross-entropy loss to ignore these positions during training.

3.8.2 Training Procedure

Training is performed in four stages.

Stage 1: Segmentation Warm-up. The Swin UNETR [1] encoder-decoder is first trained for tumor segmentation. Only the segmentation branch is optimized during this stage. The loss is given by:

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}}. \tag{3.21}$$

This stage is trained for approximately 50 epochs with early stopping based on mean Dice value of the segmentation of validation set.

Stage 2: Clinical Head Training. The laterality and anatomy heads are then trained. During this stage, only the Swin UNETR [1] backbone, laterality head, and anatomy head are updated, while the remaining modules remain frozen. The loss function is

$$\mathcal{L} = \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{lat}}\mathcal{L}_{\text{lat}} + \lambda_{\text{anat}}\mathcal{L}_{\text{anat}}. \tag{3.22}$$

This stage allows the model to learn clinically meaningful location information that will

later be used as hints for report generation. This stage is trained for around 50 epochs with early stopping based on validation loss.

Stage 3: Report Alignment. Next, the language model is trained while all image-processing modules remain frozen. Low-Rank Adaptation (LoRA) is used with rank $r = 16$ and scaling factor $\alpha = 32$. The loss function is the standard causal language-modeling loss:

$$\mathcal{L} = \lambda_{\text{LLM}} \mathcal{L}_{\text{LLM}}. \tag{3.23}$$

This stage is also trained for around 50 epochs with early stopping based on validation loss.

Stage 4: Joint Refinement. Finally, the segmentation and report-generation branches are jointly optimized through the iterative refinement module. During this stage, the Swin UNETR [1] backbone, laterality head, anatomy head, cross-attention refinement module, and language model are trained.

The overall loss is given by :

$$\mathcal{L} = \lambda_{\text{seg}} \mathcal{L}_{\text{seg}} + \lambda_{\text{lat}} \mathcal{L}_{\text{lat}} + \lambda_{\text{anat}} \mathcal{L}_{\text{anat}} + \lambda_{\text{LLM}} \mathcal{L}_{\text{LLM}}. \tag{3.24}$$

3.8.3 Training and Inference

The refinement process differs between training and inference.

During training, teacher forcing is employed. After the initial segmentation and report-generation steps, the ground-truth report tokens are used as input to the refinement module regardless of the report generated by the model. This ensures stable optimization and prevents error accumulation during training.

During inference, the report generated by the model in the first pass is used to create the refinement features. These predicted report features are then fed back into the cross-attention refinement module to update the image bottleneck representation. The refined bottleneck is subsequently used by the segmentation decoder and report-generation branch in the next refinement iteration.

Chapter 4

Experimental Details

4.1 Dataset

We use the TextBraTS [4] dataset, a volume-level text-image brain tumor dataset derived from the BraTS2020 [21, 22, 23] training set, for the experiments. This dataset contains 369 brain tumor MRI cases, where each case includes four MRI modalities: T1, T1C, T2, and FLAIR, along with a corresponding radiology report. We get the ground truth segmentation masks from the BraTS2020 [21, 22, 23] dataset. The segmentation annotations focus on three clinically meaningful tumor regions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET).

Of the 369 text-image pairs, we use 221 for training, 55 for validation and 93 for testing.

4.2 Training Setup

Most of the experiments are conducted on RTX A4000 GPU using PyTorch, MONAI and Unsloth. We resize the images into $96 \times 96 \times 96$ volume for the experiments and ablation studies. The final comparison with other models are done using images resized to $128 \times 128 \times 128$ on an A30 GPU to match the image size reported in papers. We also normalize the images channel-wise using only non-zero voxel. We use Swin UNETR [1] as image segmentation backbone and Llama-3.2-1B Instruct [24] as causal language model (LLM) for text generation. We use bf16 mixed precision with batch size of 1 and gradient accumulation of 16. We also use LoRA with r = 16 and $\alpha = 32$. We use AdamW as optimizer and use a stage-wise optimization strategy. For each training stage, only the modules required for that stage are unfrozen, while the remaining modules are kept frozen.

The Swin UNETR [1] backbone is trained with learning rate lr_{swin}, the lesion tokenizer with $\text{lr}_{\text{tokenizer}}$, the LoRA parameters of the report decoder with lr_{LoRA}, and the refinement and clinical heads with $\text{lr}_{\text{refine}}$. In the joint refinement stage, smaller learning rates are used for previously trained modules by multiplying all learning rates by 0.2. We use $\text{lr}_{\text{swin}} = \text{lr}_{\text{tokenizer}} = \text{lr}_{\text{LoRA}} = 4e-4$ and $\text{lr}_{\text{refine}} = 1e-5$. All the stages are trained for 50 epochs with early stopping with a patience of 5.

For report generation during inference, we use nucleus sampling with top-$p = 0.9$ and temperature 0.7

4.3 Evaluation Metrics

4.3.1 Segmentation Metrics

We mainly compare the Dice and HD95 (95th percentile Hausdorff Distance) for each of the sub-regions as well as the mean Dice and HD95.

Dice score measures the overlap between the predicted segmentation mask and the ground-truth mask. The higher the Dice score, the better the segmentation.

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}, \tag{4.1}$$

where P denotes the predicted mask and G denotes the ground-truth mask.

On the other hand, HD95 measures the boundary distance between the predicted mask and the ground-truth mask. Lower the HD95, the better the segmentation.

4.3.2 Report Generation Metrics

We compare BLEU (BLEU 1 & BLEU 4), ROUGE (ROUGE1 & ROUGE2 & ROUGEL), METEOR, CIDEr and BERTScore (BERTScore-p & BERTScore-r & BERTScore-f1).

BLEU measures the word overlap precision between the generated and reference reports. BLEU1 measures the single word overlap while BLEU4 measures the overlap of 4 word phrases. The higher the score, the better the report generation.

ROUGE measures word overlap recall between the generated and reference reports. ROUGE1 and ROUGE2 measures the overlap of 1-gram and 2-gram between the generated and reference reports, while ROUGEL uses the longest common sub-sequence to

measure similarity. The higher the ROUGE score, the better the report generation.

METEOR evaluates similarity using word matching, stemming, synonym matching, and word order penalty. The higher the METEOR score, the better the report generation.

CIDEr measures consensus between generated and reference reports using TF-IDF-weighted n-gram similarity. A higher CIDEr score indicates that the generated report is more consistent with the ground-truth report.

BERTScore measures semantic similarity between the generated and reference reports using contextual embeddings from transformer models like BERT. BERTScore-p, BERTScore-r and BERTScore-f1 measures the semantic precision, recall and F1 respectively. A higher BERTScore indicates that the generated report is semantically more similar to the reference report.

4.3.3 Clinical Grounding Metrics

We mainly compare the clinical F1 metrics for laterality, anatomy and pathology. We use the metrics defined in Brain3D [20] for comparison.

The Clinical Laterality F1 measures whether the report correctly identifies tumor side: left, right, or bilateral. A higher score represents the model is able to better identify the tumor side.

The Clinical Anatomy F1 measures whether the report correctly mentions involved anatomical regions, such as frontal, parietal, temporal, occipital, cerebellum and ventricle. A higher score represents the model is able to better identify the anatomical region the tumor is located.

The Clinical Pathology F1 measures whether the report correctly describes pathology-related terms, such as tumor, edema, necrosis, enhancing region, or lesion appearance. A higher score indicates the model is able to generate correct pathology-related concepts.

4.4 Ablation Study

We conduct ablation studies to analyze the contribution of different components of the model. We check the outputs after the final joint refinement stage. All ablations are conducted on input image of size $96 \times 96 \times 96$

4.4.1 Effect of Iterative Refinement

For this ablation study, we did not use any cross/iterative refinement. While the segmentation showed slightly better results without the refinement, the clinical laterality and anatomy metrics performed much better with refinement.

The iterative refinement module enables bidirectional information exchange between the segmentation and report-generation branches. In this process, the segmentation branch is refined using report-derived text features, while the updated segmentation output provides improved visual evidence for subsequent report generation. Therefore, the final generated reports become more consistent with the predicted tumor regions, leading to improved clinical grounding performance, particularly in laterality and anatomical localization.

Table 4.1: Segmentation metrics comparison.

Variant	Dice TC	Dice WT	Dice ET	Dice Mean	HD95 TC	HD95 WT	HD95 ET	HD95 Mean
Without Refinement	**0.7888**	**0.8891**	**0.7627**	**0.8125**	5.4571	**6.0723**	**4.0963**	**4.9700**
Baseline	0.7858	0.8883	**0.7627**	0.8111	**5.4405**	6.2380	4.1517	5.0334

Table 4.2: Report-generation metrics comparison.

Variant	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT-P	BERT-R	BERT-F1	CIDEr
Without Refinement	**0.5163**	0.1990	**0.6582**	0.3488	**0.4844**	**0.4040**	**0.9231**	**0.9198**	**0.9214**	0.1565
Baseline	0.5092	**0.2002**	0.6454	**0.3507**	0.4808	0.3902	0.9213	0.9179	0.9195	**0.1922**

Table 4.3: Clinical entity-level metrics comparison.

Variant	Laterality F1	Anatomy F1	Pathology F1
Without Refinement	0.7796	0.7441	**0.9976**
Baseline	**0.8029**	**0.8144**	**0.9976**

4.4.2 Effect of Clinical Grounding Hints

We check whether the two anatomy and laterality heads are required. We remove the two heads as well as the clinical hints for ablation. We observe the report and clinical metrics to be much higher on being trained with hints derived from the two heads.

The clinical hints provide important information about the tumor laterality and anatomical region. This additional guidance helps the language model to generate reports that

are more consistent with the actual tumor location. As a result, the model achieves improved report-generation metrics and substantially better clinical grounding performance, particularly for laterality and anatomy prediction.

Table 4.4: Segmentation metrics comparison for the prompt ablation study.

Variant	Dice TC	Dice WT	Dice ET	Dice Mean	HD95 TC	HD95 WT	HD95 ET	HD95 Mean
Baseline	0.7858	**0.8883**	**0.7627**	**0.8111**	**5.4405**	6.2380	**4.1517**	**5.0334**
No Hints	**0.7889**	0.8869	0.7581	0.8102	5.9143	**5.9235**	4.7324	5.2889

Table 4.5: Report-generation metrics comparison for the prompt ablation study.

Variant	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT-P	BERT-R	BERT-F1	CIDEr
Baseline	**0.5092**	**0.2002**	**0.6454**	**0.3507**	**0.4808**	**0.3902**	**0.9213**	**0.9179**	**0.9195**	**0.1922**
No Hints	0.4618	0.1418	0.5905	0.2773	0.4115	0.3444	0.9059	0.9037	0.9048	0.1212

Table 4.6: Clinical entity-level metrics comparison for the prompt ablation study.

Variant	Laterality F1	Anatomy F1	Pathology F1
Baseline	**0.8029**	**0.8144**	**0.9976**
No Hints	0.5161	0.4853	0.8847

4.4.3 Effect of Geometry Tokens and Anatomy Tokens

We check if the geometry and the anatomy tokens are needed for report generation or not. We mainly concentrate upon the report generation metrics that shows the highest value when both the token groups are present.

The geometry tokens encode explicit spatial information, including the centroid coordinates and volume of tumor-related regions such as edema, necrosis, and enhancing tumor. In contrast, the anatomy tokens provide learnable queries that capture global anatomical context from the visual memory. Together, these token groups help the language model better understand the anatomical structure of the brain as well as the location and size of the tumor regions.

Table 4.7: Segmentation metrics comparison for different ablation settings using joint refinement.

Variant	Dice TC	Dice WT	Dice ET	Dice Mean	HD95 TC	HD95 WT	HD95 ET	HD95 Mean
Baseline	0.7858	0.8883	0.7627	0.8111	5.4405	**6.2380**	4.1517	5.0334
WGWA	**0.7897**	0.8888	0.7640	**0.8128**	5.3196	6.3492	3.9476	4.9413
WG	**0.7897**	0.8888	**0.7641**	**0.8128**	5.3181	6.3393	**3.9444**	**4.9357**
WA	0.7896	**0.8889**	0.7640	**0.8128**	**5.3177**	6.3366	3.9557	4.9385

Table 4.8: Report-generation metrics comparison for different ablation settings using joint refinement.

Variant	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT-P	BERT-R	BERT-F1	CIDEr
Baseline	**0.5092**	**0.2002**	**0.6454**	**0.3507**	**0.4808**	**0.3902**	0.9213	0.9179	0.9195	**0.1922**
WGWA	0.4919	0.1682	0.6356	0.3162	0.4583	0.3740	0.9219	0.9163	0.9190	0.1726
WG	0.4956	0.1783	0.6330	0.3244	0.4656	0.3800	0.9205	**0.9181**	0.9192	0.1532
WA	0.4982	0.1758	0.6342	0.3239	0.4672	0.3750	**0.9225**	0.9171	**0.9198**	0.1676

Table 4.9: Clinical entity-level metrics comparison for different ablation settings using joint refinement.

Variant	Laterality F1	Anatomy F1	Pathology F1
Baseline	**0.8029**	**0.8144**	**0.9976**
WGWA	0.4875	0.7586	**0.9976**
WG	0.4875	0.7607	**0.9976**
WA	0.5125	0.7580	**0.9976**

Legend: Baseline = model with both geometry features and anatomy tokens; WG = without geometry features; WA = without anatomy tokens; WGWA = without both geometry features and anatomy tokens.

4.4.4 Effect of Input Modality

We check the importance of each of the modalities. We mainly check the segmentation metrics that shows the highest performance for baseline model using all the modalities. We also observe a sharp rise in clinical anatomy metric when using all the modalities.

Each of the four MRI modalities provides complementary information for segmenting different tumor regions. T1 and T1C are useful for anatomical structures and enhancement, while T2 and FLAIR are useful for edema and tumor spread. Thus all the four modalities together gives one the best segmentation results.

Table 4.10: Segmentation metrics comparison for modality ablation using joint refinement.

Variant	Dice TC	Dice WT	Dice ET	Dice Mean	HD95 TC	HD95 WT	HD95 ET	HD95 Mean
Baseline	**0.7858**	**0.8883**	**0.7627**	**0.8111**	**5.4405**	**6.2380**	**4.1517**	**5.0334**
FLAIR	0.5689	0.8647	0.3706	0.6100	10.0180	6.9314	10.2874	9.1148
T1C	0.3076	0.4094	0.2288	0.3182	28.4277	31.5405	26.2745	29.2031
T1	0.4374	0.6918	0.2569	0.4697	14.0971	13.1198	13.9967	13.3426
T2	0.4863	0.8146	0.2912	0.5383	11.4580	7.0678	11.3104	9.8001

Table 4.11: Report-generation metrics comparison for modality ablation using joint refinement.

Variant	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT-P	BERT-R	BERT-F1	CIDEr
Baseline	0.5092	**0.2002**	0.6454	0.3507	0.4808	0.3902	0.9213	0.9179	0.9195	**0.1922**
FLAIR	0.5042	0.1927	0.6465	0.3457	0.4712	0.3958	0.9185	0.9158	0.9170	0.1284
T1C	**0.5123**	0.1958	**0.6522**	**0.3517**	**0.4820**	0.3924	0.9214	0.9157	0.9185	0.1625
T1	0.5021	0.1967	0.6446	0.3474	0.4809	0.3949	**0.9252**	0.9189	**0.9220**	0.1379
T2	0.4992	0.1942	0.6455	0.3402	0.4811	**0.3981**	0.9243	**0.9196**	0.9219	0.1786

Table 4.12: Clinical entity-level metrics comparison for modality ablation using joint refinement.

Variant	Laterality F1	Anatomy F1	Pathology F1
Baseline	0.8029	**0.8144**	**0.9976**
FLAIR	**0.8495**	0.7528	0.9945
T1C	0.8315	0.7556	**0.9976**
T1	0.8136	0.7616	**0.9976**
T2	0.8136	0.7676	**0.9976**

Legend: Baseline = full four-modality input using T1, T1C, T2, and FLAIR; FLAIR = model trained and tested using only the FLAIR modality; T1C = model trained and tested using only the contrast-enhanced T1 modality; T1 = model trained and tested using only the T1 modality; T2 = model trained and tested using only the T2 modality.

4.5 Comparative Performance Analysis

We compare the proposed method with several state of the art models. All baseline comparison are done using $128 \times 128 \times 128$ images.

Segmentation-only baseline : We compare the segmentation with 3D U-Net [7], nnU-Net [3], Swin UNETR [1] and TextBraTS [4]. Compared with the state of the art

segmentation models, the model obtains a lower Dice scores, especially for ET and TC. The model performs relatively well for WT Dice, where joint refinement achieves a Dice score of 89.02, which is higher than 3D U-Net [7] and nnU-Net [3] but still lower than Swin UNETR [1] and TextBraTS [4] result.

Overall, the results show that joint refinement slightly improves the segmentation quality over report alignment, both in average Dice and average HD95, but the model still lags behind state of the art methods in segmentation.

Table 4.13: Segmentation comparison with existing methods. Higher Dice is better, and lower HD95 is better.

Method	Dice ↑				HD95 ↓			
	ET	WT	TC	Avg.	ET	WT	TC	Avg.
3D U-Net	0.804	0.873	0.816	0.831	6.11	10.51	8.93	8.17
nnU-Net	0.822	0.875	0.826	0.841	**4.27**	11.90	8.52	8.23
swin UNETR	0.810	0.895	0.808	0.838	5.95	8.23	7.03	7.07
TextBraTS	**0.833**	**0.899**	**0.828**	**0.853**	4.58	**5.48**	**5.34**	**5.13**
Proposed Work: RA	0.7692	0.8891	0.7882	0.8140	5.43	7.52	7.55	6.32
Proposed Work: JR	0.7683	0.8902	0.7932	0.8160	5.22	7.66	7.28	6.21

Legend: RA = Report Alignment; JR = Joint Refinement.

Report-generation-only baseline :

For report generation, we compare proposed method with Med3DVLM [5], MedGemma [6], and Brain3D [20]. MedGemma [6] supports 2D image inputs; therefore, for this baseline, we uniformly sample 64 axial slices of size (128×128) from the FLAIR modality for report generation. This baseline allows us to study the effect of slice-based MRI report generation compared with proposed 3D volume-based report generation framework. For Med3DVLM [5] and Brain3D [20], we use the $(128 \times 128 \times 128)$ FLAIR volume as input, since the released baseline pipelines do not natively support the same four-channel BraTS input format used by the proposed model, which consists of T1, T1C, T2, and FLAIR. Therefore, FLAIR is used as the common input modality for the external VLM baselines to ensure input-format compatibility.

Furthermore, we use the following prompt for MedGemma [6] and Med3DVLM [5] during report generation: *Generate a radiology report for this brain MRI FLAIR scan. Focus only on tumor-related findings. Describe laterality, anatomical location, edema, necrosis, enhancement if visible, and mass effect.*

The proposed model clearly outperforms all external baselines. Joint Refinement (JR)

gives the best overall report-generation performance, achieving the highest BLEU-1, BLEU-4, ROUGE-1, ROUGE-2, METEOR, BERT-F1, and CIDEr. Report Alignment (RA) gives the best ROUGE-L and slightly better anatomy F1. Clinically, both RA and JR are much stronger than the other three models, especially for laterality and anatomy grounding.

Table 4.14: Report-generation metric comparison on the tumor-only TextBraTS test set. RA denotes Report Alignment, and JR denotes Joint Refinement. Higher values indicate better performance.

Model	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT-F1	CIDEr
Med3DVLM	0.1192	0.0096	0.2500	0.0552	0.1769	0.1153	0.8540	0.0009
MedGemma	0.2631	0.0378	0.3314	0.0907	0.2123	0.1789	0.8524	0.0161
Brain3D	0.3264	0.0949	0.4409	0.1626	0.2823	0.2588	0.8928	0.0742
Proposed work: RA (FLAIR-only)	0.4785	0.1866	0.6455	0.3428	0.4972	0.3763	0.9223	0.1390
Proposed work: JR (FLAIR-only)	0.5015	0.1932	0.6457	0.3462	0.4785	0.3815	0.9190	**0.1533**
Proposed work: RA	0.4875	0.1971	0.6423	0.3484	**0.5017**	0.3795	0.9217	0.1434
Proposed work: JR	**0.5163**	**0.2038**	**0.6532**	**0.3505**	0.4833	**0.4020**	**0.9226**	0.1450

Table 4.15: Clinical entity-level metric comparison on the tumor-only TextBraTS test set. RA denotes Report Alignment, and JR denotes Joint Refinement. Higher values indicate better performance.

Model	Laterality F1	Anatomy F1	Pathology F1
Med3DVLM	0.2771	0.3335	0.2756
MedGemma	0.3118	0.4261	0.2082
Brain3D	0.5789	0.5537	0.9471
Proposed work: RA (FLAIR-only)	0.8100	0.7457	**0.9976**
Proposed work: JR (FLAIR-only)	0.8190	0.7544	0.9945
Proposed work: RA	0.8136	**0.7607**	**0.9976**
Proposed work: JR	**0.8459**	0.7539	**0.9976**

RA denotes Report Alignment, and JR denotes Joint Refinement. The best value in each column is shown in bold.

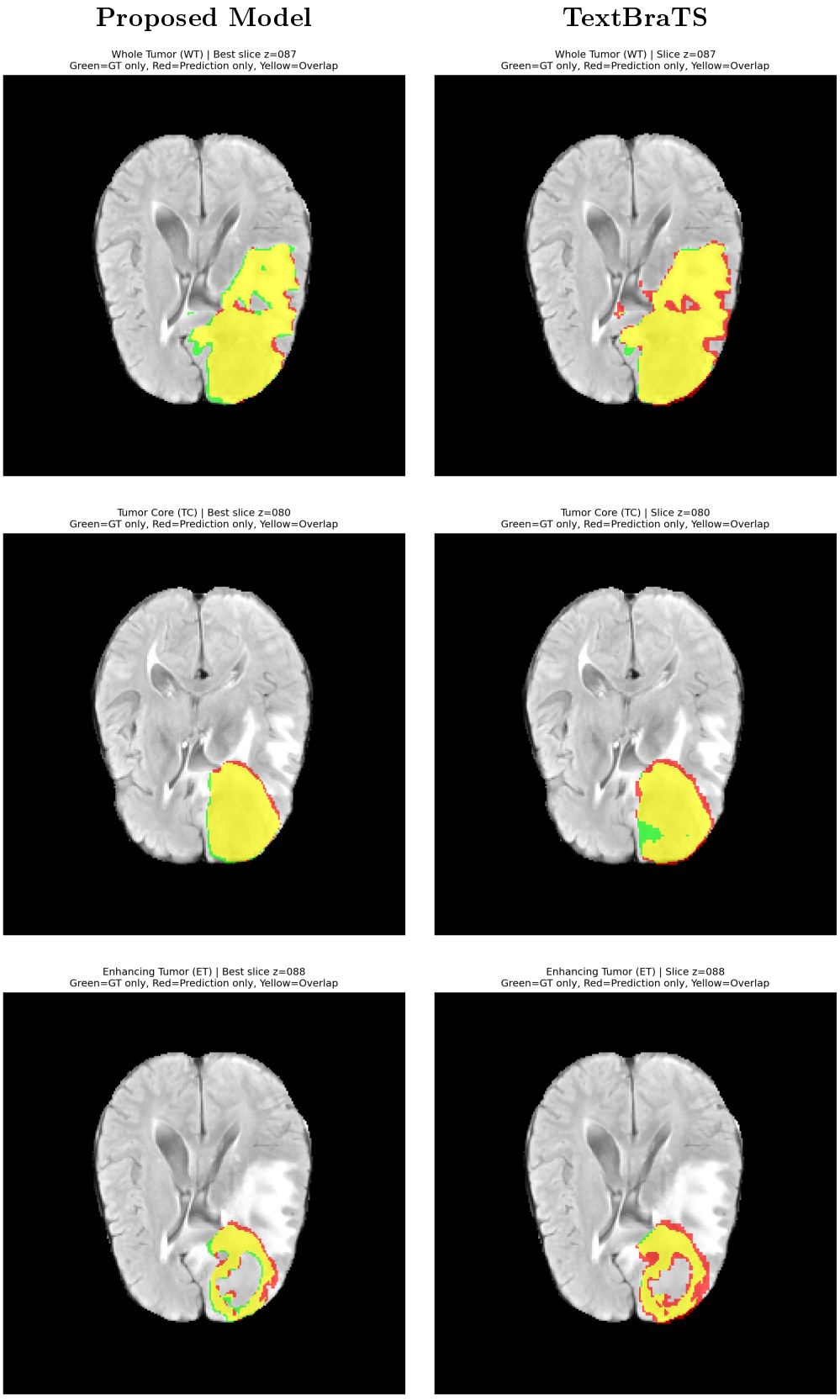

Figure 4.1: Comparison between the proposed model and TextBraTS model. The left column shows the output of the proposed model, while the right column shows the corresponding TextBraTS output. False-positive regions are shown in red, false-negative regions in green, and true-positive regions in yellow.

Table 4.16: Qualitative comparison of generated reports for the case shown in Figure 4.1

Qualitative Comparison of Generated Reports	
Ground Truth:	The lesion area is in the left frontal lobe and partially in the parietal lobe, characterized by mixed signals of varying intensity and spotty high signal areas. Edema is distributed in the left temporal and parietal lobes, extending from the left frontoparietal region to the left temporal lobe, covering a large area. Necrosis is in the central region of the lesion, particularly in the left parietal lobe, with low signal intensity and some high signal spots. Ventricular compression is observed as the left lateral ventricle is compressed and deformed.
Proposed Model:	The lesion area is in the left frontal and parietal lobes with a mixture of heterogeneous high and low signals, with spotty high signal areas suggesting multiple lesions. Edema is significant, surrounding the lesion areas, particularly in the left frontal and parietal lobes, and covers a large area of heterogeneous signal intensity, indicating significant swelling in the tissue surrounding the lesions. Necrosis is suggested by mixed high and low signal areas, with concentrated signals in certain regions, and some areas show extremely low signals, indicating possible tissue necrosis. Ventricular compression is observed with the left lateral ventricle compressed and deformed, showing slight compression.
Brain3D:	The lesion area is in the left frontal lobe with mixed high and low signals, including patchy areas of uneven signal intensity. Edema is significant around these lesions, mainly concentrated near the edges but also extending to surrounding tissues like parts of the temporal lobes as well. Necrosis may be present within some regions characterized by very dark images or a lack of discernible tissue structures due to severe destruction of normal brain matter at certain points where necrosis might occur. Ventricular compression occurs on both sides from pressure caused by edema; it shows signs of deformation compared to adjacent ventricles that are relatively unchanged.
Med3DVLM:	The mass effect appears in the image. It is a right-sided lesion located in the anatomical location of the temporal lobe. The mass effect is due to an enhancing tumor. Edema is present. The enhancement is enhancing. The anatomical location is an anatomical location in the right hemisphere. The lateralization is right.
MedGemma:	FINDINGS: There is a heterogeneously enhancing mass in the left frontal lobe measuring 2.8 x 2.5 x 2.3 cm. There is associated vasogenic edema in the left frontal lobe. There is no significant mass effect on the adjacent brain parenchyma. There is no midline shift. There is no evidence of hemorrhage. There is no hydrocephalus. IMPRESSION: Left frontal lobe enhancing mass with vasogenic edema.

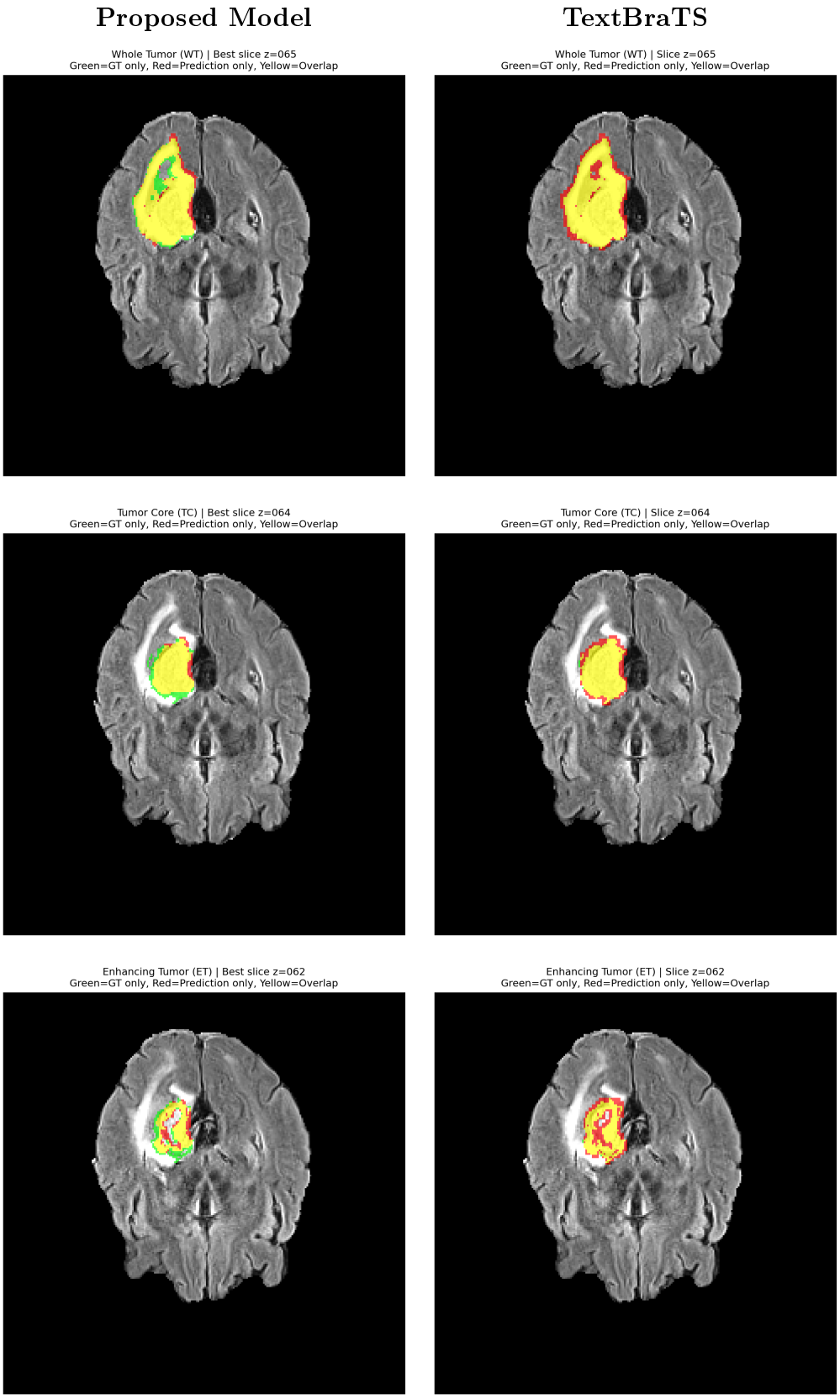

Figure 4.2: Comparison between the proposed model and TextBraTS model. The left column shows the output of the proposed model, while the right column shows the corresponding TextBraTS output. False-positive regions are shown in red, false-negative regions in green, and true-positive regions in yellow.

Table 4.17: Qualitative comparison of generated reports for the case shown in Figure 4.2

Qualitative Comparison of Generated Reports	
Ground Truth:	The lesion area is in the right frontal and occipital lobes with mixed signals of varying intensities and speckled high-signal areas. Edema appears as significant high signals, suggesting prominent edema mainly concentrated in the right occipital lobe, extending widely and encompassing most of the surrounding tissue of the lesion areas. Necrosis appears as low-signal areas observed in the right frontal and occipital lobes, with necrotic areas scattered among the high-signal lesion regions and relatively concentrated in the core of the lesions. Ventricular compression is observed, with some ventricles compressed and slightly deformed, although no significant ventricular obstruction is observed.
Proposed Model:	The lesion area is in the right frontal and parietal lobes with a mixture of heterogeneous high and low signals, along with speckled high-signal areas. Edema is significant, primarily observed in the right frontal and parietal lobes, with surrounding tissue swelling that could compress the nearby normal brain tissue. Necrosis appears as low-signal areas, suggesting possible necrosis primarily concentrated in the right frontal and parietal lobes. The necrotic areas are relatively concentrated and mainly appear as mixed signals. Ventricular compression is observed, particularly in the right lateral ventricle, which is compressed and deformed, indicating a significant compressive effect of the lesions on the brain tissue and ventricles.
Brain3D:	The lesion area is in the right frontal lobe and part of the parietal region with a mixture of heterogeneous high-signal areas along with speckled regions. Edema mainly appears as uniform high signals, indicating significant swelling around parts of these lesions. It covers most of the affected brain tissue outside normal structures but does not show any clear demarcation from surrounding healthy tissues or distinct boundaries between different edema zones. Necrosis appears primarily within low-signal mixed patches that are scattered throughout some higher-signal necrotic sections, suggesting uneven distribution concentrated at deeper levels rather than being uniformly distributed across all necrosis sites. Ventricular compression occurs due to pressure exerted on the ventricular systems, showing slight deformation while maintaining relatively stable shapes without obvious signs of severe obstruction.
Med3DVLM:	No other abnormalities were found.
MedGemma:	**FINDINGS:** There is a large, lobulated, heterogeneously enhancing mass centered in the right frontal lobe, extending into the right temporal lobe and right parietal lobe. The mass is associated with significant vasogenic edema in the right frontal, temporal, and parietal lobes. There is no significant mass effect on the adjacent brain parenchyma. There is no midline shift. There is no evidence of hemorrhage. **IMPRESSION:** Large, enhancing mass in the right frontal lobe with significant vasogenic edema.

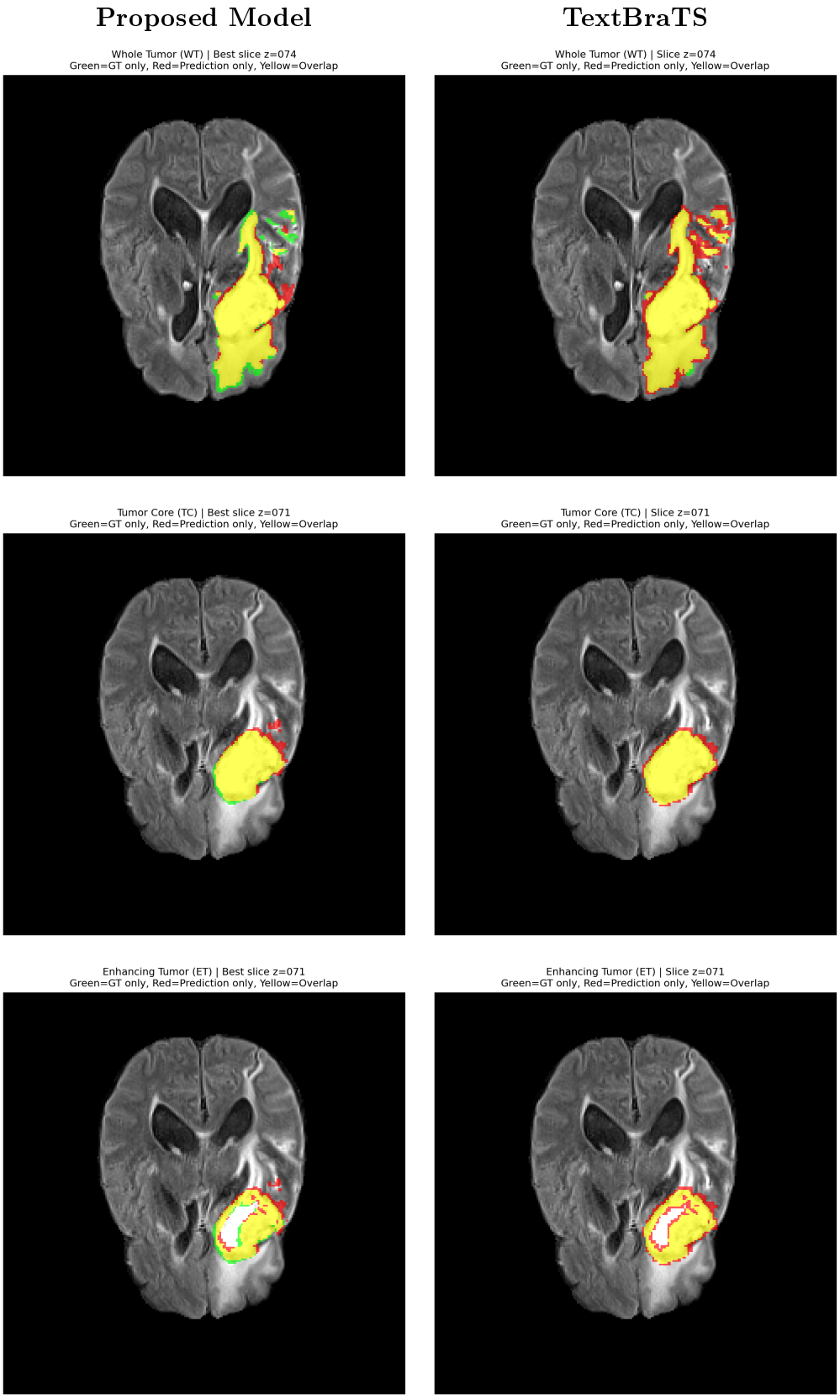

Figure 4.3: Comparison between the proposed model and TextBraTS model. The left column shows the output of the proposed model, while the right column shows the corresponding TextBraTS output. False-positive regions are shown in red, false-negative regions in green, and true-positive regions in yellow.

Table 4.18: Qualitative comparison of generated reports for the case shown in Figure 4.3

Qualitative Comparison of Generated Reports	
Ground Truth:	The lesion area is in the left frontal and parietal lobes with heterogeneous signal intensities, displaying both high and low signal areas, along with speckled high-signal regions. Edema is evident around the lesions in the left frontal and parietal lobes, primarily observed around the high-signal lesions, indicating the extent of tissue swelling and edema. Necrosis is mainly concentrated in the left parietal region, characterized by low signal intensity, with some areas showing mixed signals and varying degrees of tissue necrosis. Ventricular compression is observed with the right lateral ventricle slightly compressed, and the shape of the left lateral ventricle somewhat altered.
Proposed Model:	The lesion area is in the left frontal and parietal lobes with mixed high and low signals with patchy high-signal areas. Edema is significant, primarily observed in the left frontal and parietal lobes, with high-signal areas indicating swelling in the tissues surrounding the lesion and affecting a large area of brain tissue. Necrosis is observed as low-signal areas within the lesion regions, mainly concentrated in the left frontal and parietal lobes, with mixed high-signal regions indicating necrosis. These necrotic areas are relatively dispersed and adjacent to edema areas. Ventricular compression is observed, with the left lateral ventricle compressed and deformed, showing slight compression.
Brain3D:	The lesion area is in the right frontal lobe with heterogeneous mixed high and low signals, including patchy areas of high signal. Edema is significant around the lesion and is mainly concentrated on the peripheral region surrounding it, appearing as a large extent of abnormal signal outside the main lesion zone. Necrosis may be present within some very dark or low-intensity regions compared to adjacent edema-affected white matter, appearing more scattered than the necrotic regions described in the ground truth. Ventricular compression occurs at both ventricles under pressure from swelling near the ventricular system, although obvious deformation is not clearly described.
Med3DVLM:	No other pathology is visible. The mass lesion is again observed in the left temporal lobe, hypodense with no enhancement, measuring 6.0 $\times$ 5.0 $\times$ 4.2 cm. The mass demonstrates mild heterogeneous enhancement in the right lateral wall measuring 10 mm thick; however, no extra-axial extension is seen and no acute intracranial hemorrhage is present. There is no edema or necrosis identified.
MedGemma:	**FINDINGS:** There is a large, heterogeneously enhancing mass centered in the right frontal lobe, extending into the right temporal lobe and involving the right lateral ventricle. There is significant vasogenic edema surrounding the mass. There is no midline shift or herniation. There is no evidence of intracranial hemorrhage. **IMPRESSION:** Large, enhancing mass in the right frontal/temporal lobes with significant vasogenic edema.

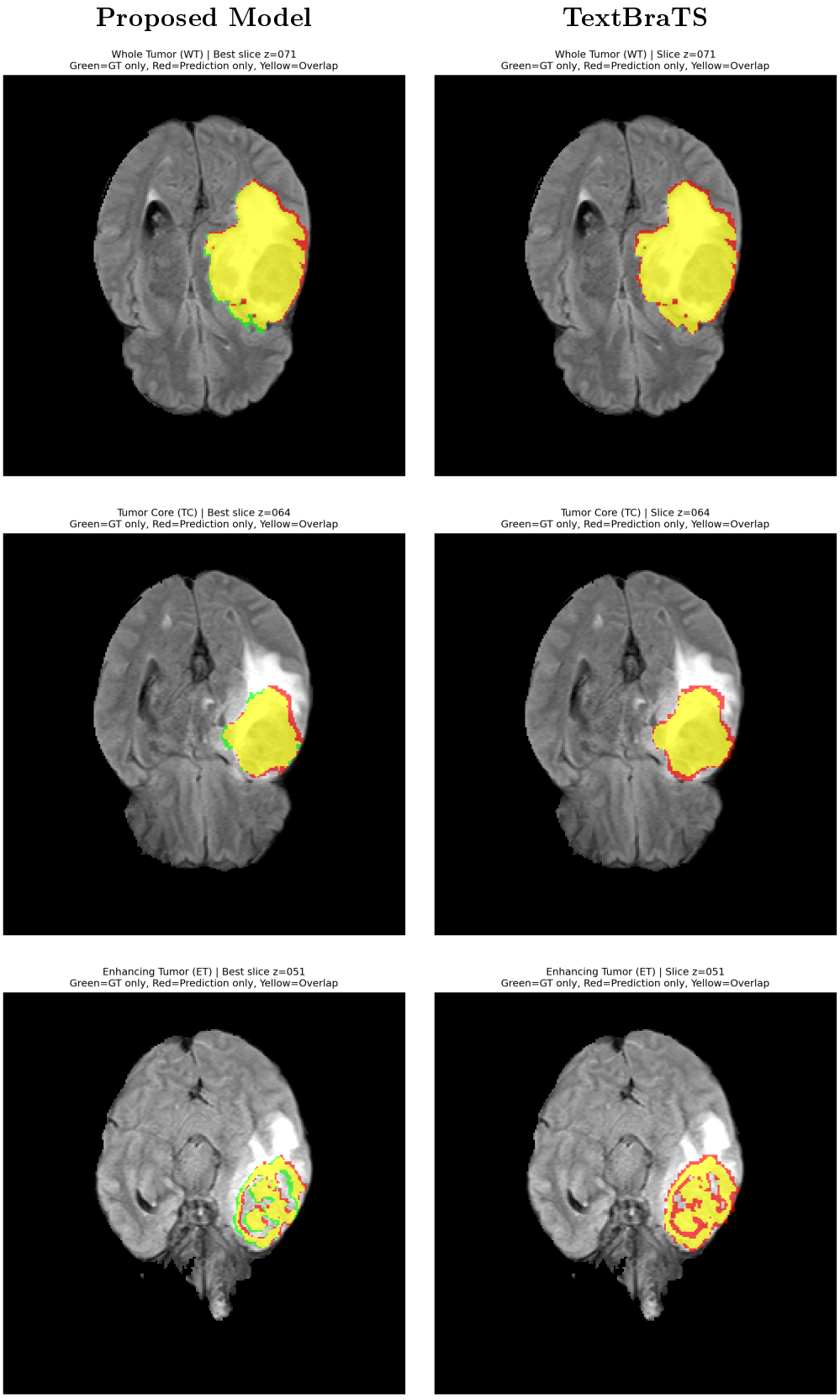

Figure 4.4: Comparison between the proposed model and TextBraTS model. The left column shows the output of the proposed model, while the right column shows the corresponding TextBraTS output. False-positive regions are shown in red, false-negative regions in green, and true-positive regions in yellow.

Table 4.19: Qualitative comparison of generated reports for the case shown in Figure 4.4

Qualitative Comparison of Generated Reports	
Ground Truth:	The lesion area is in the left frontal and parietal lobes with a mixture of high and low signal spots with patchy high signals. Edema is significant, observed around the lesion areas in the left frontal and parietal lobes, especially within and surrounding the lesions, characterized by high-intensity signals. Necrosis is observed as low-signal areas interspersed within the high-signal regions, particularly in the central portion of the lesions with a mixture of high and low signals. Ventricular compression is mild.
Proposed Model:	The lesion area is in the left frontal and parietal lobes with a mixture of heterogeneous high and low signals, with some areas showing speckled high signals. Edema is significant, surrounding the lesion areas in the left frontal and parietal lobes, with marked swelling of the tissues around these regions and a considerable extent of edema. Necrosis is present within the lesion areas, with regions within the left frontal and parietal lobes displaying low signal intensity, suggesting possible necrotic tissue. These necrotic areas are relatively concentrated. Ventricular compression is observed, with the left lateral ventricle being compressed and deformed.
Brain3D:	The lesion area is in the right frontal lobe with mixed high and low signals, along with speckled areas of uneven signal intensity. Edema surrounding the lesion is mainly concentrated around these regions and shows significant swelling affecting parts of the white matter tissue at the edge of a large region. Necrosis is described as scattered necrotic foci that are relatively isolated from each other, appearing mostly as irregular patches interspersed among healthy tissue sections rather than forming extensive contiguous zones.
Med3DVLM:	The mass is right-sided and located in the frontal lobe. There is no edema or necrosis. The lesion enhances heterogeneously and causes mass effect with midline shift to the left.
MedGemma:	**FINDINGS:** There is a large, heterogeneously enhancing mass centered in the right frontal lobe, extending into the right temporal lobe and right parietal lobe. The mass is associated with significant vasogenic edema in the right frontal, temporal, and parietal lobes. There is no significant mass effect on the adjacent brain parenchyma. There is no midline shift or hydrocephalus. **IMPRESSION:** Large, heterogeneously enhancing mass in the right frontal/temporal/parietal lobes with vasogenic edema.

Chapter 5

Conclusion and Further Work

In this work, we proposed a unified 3D vision-language framework for simultaneous brain tumor segmentation and radiology report generation from multi-modal MRI volumes.

The experimental results show that the proposed model is effective for both segmentation and report generation. While the model was unable to beat the state of the art methods, it still achieves strong segmentation performance, with a mean Dice score of 81.60% and an HD95 of 6.21.

For report generation, it achieves a BERTScore-F1 of 0.9226, clinical laterality F1 of 0.8459, clinical anatomy F1 of 0.7539, and clinical pathology F1 of 0.9976, indicating that the model can generate clinically meaningful reports while accurately localizing tumor regions and producing textual descriptions that are consistent with the segmented tumor areas. Moreover, through the ablation studies, we also showed that the presence of laterality and anatomy heads, geometry extraction and anatomical queries improve the laterality and anatomy F1 allowing the model to accurately locate the tumor position for report generation.

Further, we proposed a basic iterative model that allows the two tasks to learn from each other, once initial segmentation and report generation has been done. This allowed the model to improve the initial segmentation and report by leveraging the knowledge of each other.

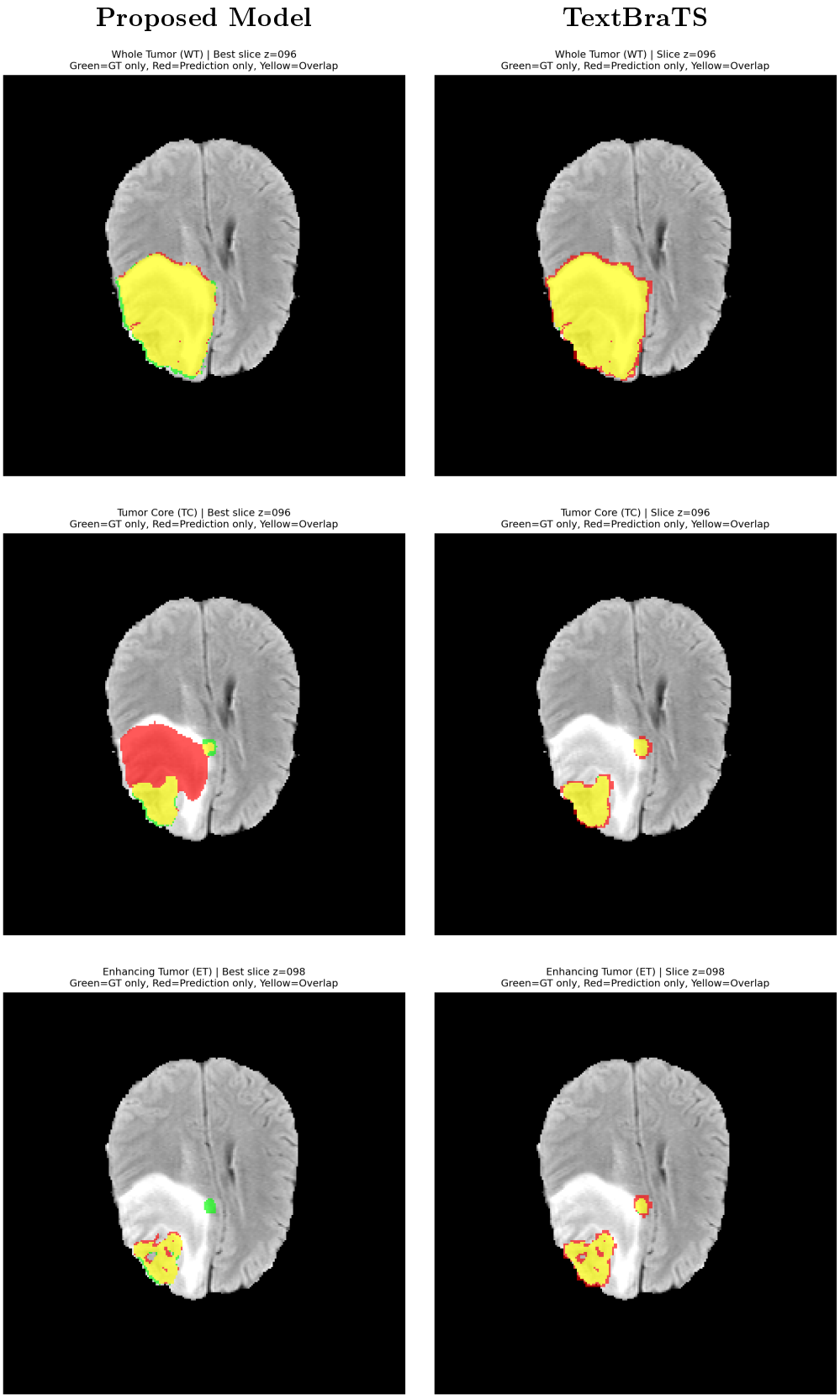

Figure 5.1: Comparison between the proposed model and TextBraTS model for a failed case. False-positive regions are shown in red, false-negative regions in green, and true-positive regions in yellow. Although the WT region segmentation is reasonably good, the TC segmentation contains many false positives, leading to a low TC Dice score.

Further Work

We are mainly looking forward to working in three directions :

1. The BraTS2020 dataset mainly contains glioma brain tumor cases. In future, we may add further brain tumor cases such as meningioma and pituitary tumors, allowing the report generation framework to identify and describe a wider range of brain tumor types. This would help improve the generalization of the model across different tumor appearances, anatomical locations, and clinical presentations.

2. Although the proposed model achieves promising report generation performance, its segmentation performance is still lower than state-of-the-art segmentation models, with a lower mean Dice score and a higher HD95 value. As shown in Figure 5.1, the TC segmentation may contain false positives, leading to a lower mean Dice score. Future work will involve investigating the factors responsible for this performance gap, such as limitations in the refinement strategy, loss formulation, input resolution, and training configuration. Furthermore, instead of using a simple cross-attention-based refinement module, we plan to explore region-based attention mechanisms, where the generated text attends separately to tokens corresponding to different tumor regions. The region-wise attended features can then be merged and used to refine the bottleneck representation, potentially improving the interaction between segmentation and report generation.

3. Although the current iterative refinement module shows promise, it can be further improved by adding a consistency-checking mechanism, either as a neural network module or as an additional loss function. This mechanism would measure the alignment between the generated text report and the segmentation output. Incorporating such a consistency constraint may lead to better refinement and improved agreement between the visual and textual outputs.

Bibliography

[1] A. Hatamizadeh, V. Nath, Y. Tang *et al.*, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, ser. Lecture Notes in Computer Science. Springer, 2022. [Online]. Available: https://doi.org/10.1007/978-3-031-08999-2_22

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, vol. 9351. Cham: Springer, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[3] F. Isensee, P. F. Jaeger, S. A. A. Kohl *et al.*, "nnU-Net: A Self-configuring Method for Deep Learning-based Biomedical Image Segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. [Online]. Available: https://doi.org/10.1038/s41592-020-01008-z

[4] X. Shi, R. K. Jain, Y. Li *et al.*, " TextBraTS: Text-Guided Volumetric Brain Tumor Segmentation with Innovative Dataset Development and Fusion Module Exploration ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, vol. LNCS 15965. Springer Nature Switzerland, September 2025.

[5] Y. Xin, G. C. Ates, K. Gong *et al.*, "Med3DVLM: An Efficient Vision-Language Model for 3D Medical Image Analysis," *IEEE Journal of Biomedical and Health Informatics*, 2025, accepted. [Online]. Available: https://arxiv.org/abs/2503.20047

[6] A. Sellergren, S. Kazemzadeh, T. Jaroensri *et al.*, "MedGemma Technical Report," *arXiv preprint arXiv:2507.05201*, 2025. [Online]. Available: https://arxiv.org/abs/2507.05201

[7] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp *et al.*, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, ser. Lecture Notes in

Computer Science, vol. 9901. Springer, 2016, pp. 424–432. [Online]. Available: https://doi.org/10.1007/978-3-319-46723-8_49

[8] Z. Liu, Y. Lin, Y. Cao *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10 002.

[9] Z. Li, Y. Li, Q. Li *et al.*, "LViT: Language Meets Vision Transformer in Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 96–107, 2024. [Online]. Available: https://doi.org/10.1109/TMI.2023.3291719

[10] T. Koleilat, H. Asgariandehkordi, H. Rivaz *et al.*, "MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, ser. Lecture Notes in Computer Science. Springer, 2024. [Online]. Available: https://doi.org/10.1007/978-3-031-72390-2_60

[11] T. Koleilat, H. Asgariandehkordi, O. N. Manzari *et al.*, "MedCLIPSeg: Probabilistic Vision-Language Adaptation for Data-Efficient and Generalizable Medical Image Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. [Online]. Available: https://github.com/HealthX-Lab/MedCLIPSeg

[12] J. Lee, W. Yoon, S. Kim *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics (Oxford, England)*, vol. 36, 09 2019.

[13] Y. Zhang, H. Jiang, Y. Miura *et al.*, "Contrastive Learning of Medical Visual Representations from Paired Images and Text," in *Proceedings of the 7th Machine Learning for Healthcare Conference*, ser. Proceedings of Machine Learning Research, Z. Lipton, R. Ranganath, M. Sendak *et al.*, Eds., vol. 182. PMLR, 05–06 Aug 2022, pp. 2–25. [Online]. Available: https://proceedings.mlr.press/v182/zhang22a.html

[14] D. A. Zifeng Wang, Zhenbang Wu and J. Sun, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022. [Online]. Available: https://aclanthology.org/2022.emnlp-main.256/

[15] L. W. L. Z. Zhanyu Wang, Lingqiao Liu, "R2GenGPT: Radiology Report Generation with Frozen LLMs," *Meta-Radiology*, vol. 1, no. 3, p. 100033, 2023. [Online]. Available: https://doi.org/10.1016/j.metrad.2023.100033

[16] M. Moor, Q. Huang, S. Wu *et al.*, "Med-Flamingo: a Multimodal Medical Few-shot Learner," in *Proceedings of the 3rd Machine Learning for Health Symposium*, ser. Proceedings of Machine Learning Research, S. Hegselmann, A. Parziale, D. Shanmugam *et al.*, Eds., vol. 225. PMLR, 10 Dec 2023, pp. 353–367. [Online]. Available: https://proceedings.mlr.press/v225/moor23a.html

[17] C. Li, C. Wong, S. Zhang *et al.*, "LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day," in *Advances in Neural Information Processing Systems 36*, 2023, datasets and Benchmarks Track. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html

[18] E. S. M. B. Hamamci, I.E., "CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, ser. Lecture Notes in Computer Science. Springer, 2024, pp. 476–486. [Online]. Available: https://doi.org/10.1007/978-3-031-72390-2_45

[19] I. E. Hamamci, S. Er, C. Wang *et al.*, "Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography," *Nature Biomedical Engineering*, 2026. [Online]. Available: https://doi.org/10.1038/s41551-025-01599-y

[20] M. Barone, F. Di Serio, G. Riccio *et al.*, "Brain3D: Brain Report Automation via Inflated Vision Transformers in 3D," 2026, arXiv preprint. [Online]. Available: https://arxiv.org/abs/2602.22098

[21] B. H. Menze, A. Jakab, S. Bauer *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[22] S. Bakas, H. Akbari, A. Sotiras *et al.*, "Advancing The Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features," *Scientific Data*, vol. 4, p. 170117, 2017.

[23] S. Bakas, M. Reyes, A. Jakab *et al.*, "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[24] Meta AI, "Llama 3.2 1B Instruct," https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct, 2024.