

MedViT-MT: A Multi-Task Medical Vision Transformer for Glioblastoma Brain Analysis with Temporal Encoding and Clinical Data Fusion

Sana Cheema¹, Ghulam Gilanie², Muhammad Faheem Mushtaq³, Rania Aboalela⁴, Ali Daud⁵

Department of Artificial Intelligence, Islamia University, Bahawalpur, Punjab 63100, Pakistan^{1, 2}

Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur, Punjab 63100, Pakistan³

Department of Information Systems-Faculty of Computing and Information Technology,
King Abdulaziz University, Rabigh, Saudi Arabia⁴

Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates⁵

Abstract—Glioblastoma Multiforme (GBM) is the most invasive form of primary brain tumor. The average survival time after diagnosis is only 15 months. Accurately and automatically analyzing GBM MRI scans remains a major clinical challenge. It requires the tumor to be localized, a report to be generated in a structured way, and information from multiple data sources to be integrated simultaneously. Existing methods address these problems separately and use models for one task that do not leverage the complementary information available across segmentation, captioning, and grounding. This study introduces MedViT-MT (Medical Vision Transformer Multi-Task), a unified transformer architecture that performs tumor segmentation, molecular-subtype classification, lesion grounding, and patient-conditioned structured caption generation from GBM DSC-MRI sequences. MedViT-MT employs Swin Transformer Tiny as its hierarchical visual transformer backbone for extracting multiscale imaging features, followed by the addition of a clinical fusion module that is based on cross-attention and that directly places tabular clinical features (IDH1, MGMT, KPS, GTR, and GTR) into the visual feature space. It further extracts inter-slice volumetric information from the MRI slice sequence, helping the model infer three-dimensional tumor volume. On the UPENN-GBM dataset, which contains 375 patients with DSC MRI sequences and their clinical records, MedViT-MT reaches 98.31% classification accuracy, 99.37% AUC-ROC, 91% mean Dice score, and 100% few-shot accuracy (1-shot, 5-shot, 10-shot). The model is also able to generalize well to different IDH1 subtypes, and it runs using 16GB RAM on a CPU. The separate experiments to identify what each architectural component delivers show that all of them matter; the statistical analyses indicate the reliability of the model.

Keywords—Glioblastoma; multi-task learning; deep learning; transformer; clinical data fusion; temporal encoding

I. INTRODUCTION

Brain tumors are one of the most severe neurological conditions that have a profound impact on the survival and quality of life of patients [1]. To support diagnosis, therapy planning, and evaluation of prognosis, it is vital to perform accurate tumor detection, precise tumor localization, and detailed characterization of tumor subregions [2]. Magnetic Resonance Imaging (MRI) is currently the most widely used imaging technique for assessing brain tumors, as it provides

excellent visualization of soft-tissue structures and supports multiple sequences [3, 4]. The use of artificial intelligence for automated brain tumor analysis has become a significant research focus driven by the increasing accessibility of multi-modal MRI datasets from sources such as The Cancer Imaging Archive and the establishment of the Brain Tumor Segmentation Challenge as a benchmark [5].

Brain tumors are usually examined through MRI scans of different sequences, such as T1, T1 contrast-enhanced (T1c), T2, and FLAIR, to identify tumor parts like the enhancing tumor, necrotic core, and edema. In the clinic, radiologists manually check these sequences to draw lesion boundaries, calculate tumor burden, and characterize radiological features through narrative reports [6]. This method takes up a lot of time, and different people may disagree on the results. As the amount of imaging data keeps increasing and there is a demand for precision medicine, automated systems that can help clinicians in tumor interpretation and decision-making are becoming necessary [7].

The clinical examination and evaluation of brain tumors nowadays is being done mainly through radiologists examining multi-sequence MRI visually using manual or semi-automatic tumor segmentation tools, grading radiologically based on morphological features to perform histopathological examination after biopsy or surgery, and integrating imaging results with clinical and molecular data. These techniques are still considered the gold standard in neuro-oncology and have a crucial role in confirming diagnosis and planning therapy [8]. However, the traditional clinical methods still have several limitations despite their great importance. Manual cut-out of the work from images is tiring and subject to differences in interpretation of the results by different experts [9]. Looking at surfaces visually may miss patterns that are located deeply in data obtained from multiple sequences. Obtaining histopathological confirmation means doing highly intrusive procedures, which cannot always be done. Clinical workflows as they stand nowadays usually generate narrative-based reports rather than reports with quantitative uncertainty estimates, which essentially means that the confidence level can only be assessed based on someone's judgment or perception. All of this

points to the necessity of dependable, automated, and interpretable computational support tools [10].

During the last several years, there has been a significant growth trend in the use of machine learning and deep learning techniques for the automated detection and segmentation of brain tumors. Convolutional Neural Networks (CNNs) such as the U-Net and its different versions have been very successful in volumetric tumor segmentation tasks [11]. Transformer-based architectures have recently taken medical imaging to a new level by enhancing the understanding of global context [12]. Classification of tumor types and prediction of survival outcomes using MRI features is also among the areas that have been extensively studied [13]. Medical image captioning is another area that is gaining attention, and it is focused on generating radiology-type descriptions from imaging data [14].

Although the majority of deep learning methods promise great results, some major issues are raised by most existing ones [15]. A large number of the models do not go beyond segmentation and classification and hence miss out on structured clinical information. The models designed for caption generation usually depend on global image features, which makes them incapable of relating textual descriptions solely to the tumor area. Besides, they usually don't take into account uncertainty estimation, which means they give overconfident predictions that may cause a loss of clinical trust [16]. Thus, at present, these tools are more like isolated predictive instruments rather than comprehensive, clinically meaningful decision-support frameworks[17].

Despite the huge steps made in brain tumor segmentation and classification, there are still quite a few gaps. First, there is not much research on single frameworks that can do segmentation, classification, and radiology-style report generation all at the same time [18]. Second, most of the vision-language methods have not been updated to include segmentation-guided grounding, so it is not guaranteed that the texts produced by these methods describe the tumor areas correctly. Third, the lack of structured clinical metadata in the model design means the potential for patient-specific reasoning is not being realized. Lastly, the importance of uncertainty-aware modeling for dependable clinical use is hardly recognized in multi-task brain tumor systems[19].

The main contributions of this research are outlined below:

- This research proposes a unified multi-task deep learning framework that simultaneously performs brain tumor segmentation, molecular-subtype classification, lesion grounding, and patient-conditioned structured caption generation, enabling comprehensive GBM analysis within a single model.
- This study introduces a segmentation-guided visual grounding mechanism that generates tumor-aware region tokens and aligns visual representations with generated structured clinical captions, thereby improving the spatial interpretability of the textual output.
- This research develops a cross-attention transformer-based fusion module to combine the structured clinical metadata, which allows the model to incorporate patient-

specific context and support more informed patient-aware decision-making.

- This study incorporates uncertainty-aware modeling to generate confidence maps with the predictions to enhance the interpretability, reliability, and trustworthiness of the system in clinical applications.

The rest of the study is structured as follows. Section II discusses the literature on brain tumor segmentation, multi-modal fusion, and medical report generation. Section III details the proposed methodology used to cover the segmentation branch, clinical fusion mechanism, uncertainty modeling, and caption generation module. Section IV is for the presentation of the quantitative results, ablation studies, and comparisons with state-of-the-art methods. Section V: Conclusion and Future Work.

II. LITERATURE REVIEW

Deep learning has made significant advancements in brain tumor segmentation and classification using MRI data. However, combining clinical data, uncertainty modeling, and report generation into a single integrated system is still quite limited.

A. Brain Tumor Segmentation in Multi-Modal MRI

A multimodal transformer-based architecture was introduced for brain tumor segmentation using incomplete MRI data [20]. The approach combines a modality-specific encoder, a multi-modal transformer, and a shared-weight decoder to generate missing modalities. The experiments on BraTS 2018, 2019, and 2020 datasets show that the model outperforms the state-of-the-art models in terms of segmentation accuracy, especially when modalities are missing. This study [21] presents a deep learning-based method for brain tumor segmentation and classification. The authors used DeepLabV3+ integrated with Atrous Spatial Pyramid Pooling (ASPP) and Bayesian Optimization. The main features of the work are a couple of custom-made models, namely, IRB-96 and IRB-Self, which are designed for feature extraction and classification, respectively. The features obtained from the model are optimized through Bayesian methods, and to make the results more understandable, LIME is used. Their method reaches 92.68% segmentation accuracy and 95.42% classification accuracy on the dataset, which is better than other advanced techniques. Recently, a novel technique, namely Contextual Transformer U-Net (CT-UNet), was introduced for brain tumor segmentation[22]. It is a hybrid of Transformer blocks and U-Net architecture that effectively outlines tumors in multimodal MRI sequences. On the BraTS dataset, CT-UNet achieved a Dice coefficient of 0.92, sensitivity of 0.90, and specificity of 0.94, therefore outperforming the conventional models. Furthermore, it showed very accurate spatial localization with a Hausdorff distance of 7.5 mm. Besides, the proposed method enables rapid and precise segmentation, which in turn allows clinical application for improving patient care and treatment planning. Multi-Modal Segment Anything Model (MSAM) for glioma tumor segmentation is proposed by integrating structural MRI with diffusion tensor imaging (DTI) data [23]. The approach feature first a multi-modal feature fusion block, missing modality training technique to handle missing data

scenarios. MSAM on the BraTS 2018 dataset outperforms U-Net not only in terms of Dice Similarity Coefficient but also in 95% Hausdorff Distance, particularly when only structural MRI data is available. These results support the ability of MSAM to discard a modality and still yield reliable segmentation, even with incomplete data or varying tumor sizes.

B. Uncertainty-Aware Segmentation for Trustworthy AI

A deep learning approach to brain tumor segmentation was proposed that combines uncertainty estimation with an attention-based multimodal fusion mechanism [24]. The presented system is based on a multi-encoder 3D U-Net architecture combined with Bayesian uncertainty estimation for accurate segmentation. The numerical comparisons on the BraTS 2018 and BraTS 2019 datasets demonstrate that their approach outperforms the state-of-the-art methods in Dice score, Hausdorff distance, and sensitivity. This method significantly improves segmentation accuracy at the tumor border, and it is flexible to different network architectures and computer vision tasks. A brain tumor segmentation method is introduced by incorporating multi-modal fusion and uncertainty quantification techniques [25]. Fuchs et al. [26] explored the feasibility of uncertainty measures in brain cancer delineation and put forward a multi-headed Variational U-Net (mu-VNet). This approach is based on the variational inference principle merged with the ensemble's variability features to produce better uncertainty assessments without sacrificing much on the computational requirement. They tested this model on the BraTS 2018 data in the in-distribution set-up and four distribution-shift scenarios. It is noteworthy that instead of the teacher-student multimodal fusion described in [25], the main novelty of [26] is reliable uncertainty estimation, calibration techniques, and resistance to out-of-distribution inputs. Such characteristics make this work especially helpful to engineers and scientists who are designing accurate systems for medical image segmentation. The FLAIR and T1c modalities instruct the T2 and T1 modalities in a teacher-student fusion manner. The model employs Monte Carlo dropout for the uncertainty calculation and introduces an uncertainty-aware loss function. On the BraTS 2018, 2019, and 2020 datasets, the method scored 86.6% Dice Similarity Coefficient for whole tumor, 87.1% for tumor core, and 79.0% for enhancing tumor, while the Hausdorff Distance was only 4.2 mm, surpassing the state-of-the-art methods.

C. Radiology Report Generation and Vision-Language Modeling

MediVLM is a proposed that is the transformer-based vision-language model that combines visual encodings and text representations through cross-attention to generate structured clinical captions from medical images [27]. Furthermore, it is capable of functioning with very limited paired training data by relying on object detection and cross-modal alignment. Another approach to automatic radiology report generation that takes into account chest X-rays [28]. It relies on basically two modules: 1)

a learned knowledge base that extracts and stores medical knowledge automatically from textual embeddings, and 2) multi-modal alignment that is used to ensure semantic congruence between X-ray images, reports, and disease labels. The proposed method was evaluated on the public IU-Xray and MIMIC-CXR datasets by means of both Natural Language Generation (NLG) and clinical efficacy (CE) metrics. The findings indicate that their model surpasses the performance of state-of-the-art methods in different metrics; thus, it proves the effectiveness of the learned knowledge base and multimodal alignment. The assessment reveals that clinical accuracy has been considerably improved, with a significant increase in precision, recall, and F1 scores, particularly in clinical efficacy metrics. Medical Vision Attention Generation (MedVAG) model is a reliable system for the automatic generation of medical reports [29]. The main idea of the model is to make use of Vision Transformer (ViT)-based visual feature extraction and combine it with GPT-2 for language modeling. The model is further improved through graph-based feature fusion and multiple attention mechanisms (co-attention, cross-attention, and memory-guided attention) that are aimed at ensuring the consistency and accuracy of the medical diagnosis. Tested on the IU X-Ray and COV-CTR datasets, MedVAG was able to outperform state-of-the-art methods in terms of several metrics, namely BLEU, METEOR, ROUGE, and CIDEr, as well as in clinical effectiveness measures. Through ablation experiments, it was found that the use of attention mechanisms and feature fusion plays a crucial role in the association of visual and textual information. MedVAG can be highly effective as a radiologist's assistant, helping to ease the workload and increase the accuracy of diagnoses.

This research identifies four critical gaps in existing literature: 1) no unified model jointly addresses segmentation, captioning, and grounding for GBM; 2) clinical tabular data is rarely fused with imaging features in a principled cross-attention framework; 3) the temporal sequential structure of DSC MRI slices is seldom exploited; and 4) few-shot and zero-shot learning modes for GBM classification remain unexplored. This work addresses all four gaps simultaneously.

III. THE PROPOSED METHODOLOGY

In this section, the complete methodology of the proposed Medical Vision Transformer Multi-Task (MedViT-MT) framework is presented, based on a deep learning unified system designed for glioblastoma multiforme (GBM) analysis using DSC-MRI sequences along with clinical metadata. The framework combines segmentation, classification, visual grounding, and patient-conditioned structured clinical caption generation into one comprehensive (end-to-end trainable) model. Unlike regular single-task models, MedViT-MT is able, via hour-by-hour cross-modal fusion and temporal reasoning, to simultaneously capture both spatial and volumetric features of tumors shown in Fig. 1.

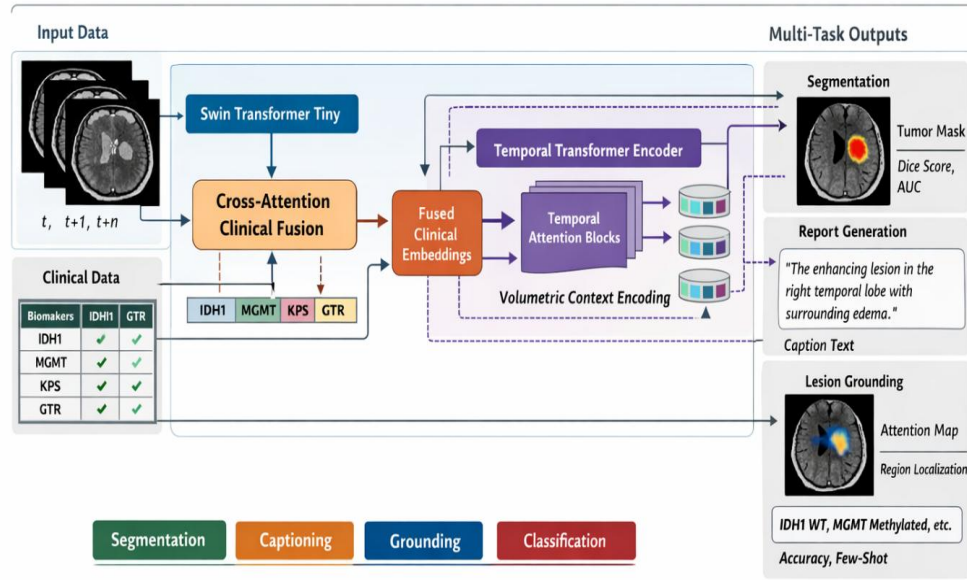


Fig. 1. MedViT-MT proposed architecture.

The input to the model consists of a sequence of MRI slices and a corresponding vector of structured clinical features. Let

$$X = \{x_1, x_2, \dots, x_T\} \quad x_T \in \mathbb{R}^{3 \times H \times W} \quad (1)$$

Denote the sequence of T DSC MRI slices, where $H=W=224$. The associated clinical vector is defined as

$$c \in \mathbb{R}^{D_c}, D_c = 7 \quad (2)$$

To optimize the four tasks segmentation, caption generation, grounding, and classification simultaneously, the model formulates a unified objective function as

$$\tau_{total} = \tau_{seg} \tau_{seg} + \tau_{cap} \tau_{cap} + \tau_{gnd} \tau_{gnd} + \tau_{cls} \tau_{cls} \quad (3)$$

where, each λ controls the contribution of its respective task. This formulation enables joint learning and cross-task regularization, improving generalization performance across all outputs.

The framework consists of four successive steps. Initially, a hierarchical transformer backbone is employed by the model to extract discriminative visual features from a series of MRI slices. Secondly, the clinical fusion module that merges the patient's structured information with the image features is used to enrich the visual features. Thirdly, a temporal transformer encoder is utilized to capture the relationships among the slices, resulting in the production of context-aware representations along the entire MRI sequence. Lastly, the representations that have been learned are forwarded to the task-specific heads, which will be responsible for producing tumor masks, lesion localization outputs, descriptive reports, and subtype predictions (Algorithm 1). The main idea behind this architecture is that the model will be able to combine pixel-level comprehension, region-level localization, patient-level classification, and language-level clinical summarization into one unified learning system.

Algorithm 1: MedViT-MT Forward Pass

```

INPUT: image_seq [B, T, 3, 224, 224], clinical [B, 7], valid_mask [B, T]
OUTPUT: seg_masks, cap_logits, boxes, scores, cls_logits
-- STAGE 1: Backbone Feature Extraction --
B, T, C, H, W = image_seq.shape
x = image_seq.reshape(B*T, C, H, W) # Flatten batch+time
img_feats = SwinBackbone(x) # [B*T, 768]
img_feats = LayerNorm(img_feats)
img_feats = Dropout(img_feats)
img_feats = img_feats.reshape(B, T, 768) # Restore time dim
-- STAGE 2: Clinical Cross-Attention Fusion --
clin_proj = MLP_clinical(clinical) # [B, 256]
clin_exp = clin_proj.unsqueeze(1).expand(B,T,256) # [B,T,256]
img_proj = Linear_img(img_feats) # [B, T, 256]
cross_attn_out, _ = MultiheadCrossAttn(Q=img_proj, K=clin_exp, V=clin_exp)
fused = LayerNorm(img_proj + Dropout(cross_attn_out))
sa_out, _ = MultiheadSelfAttn(fused, fused, fused)
fused = LayerNorm(fused + Dropout(sa_out))
fused = LayerNorm(fused + Dropout(FFN(fused))) # [B, T, 256]
-- STAGE 3: Temporal Encoder --
cls_token = learnable_cls.expand(B, 1, 256)
x_temp = concat([cls_token, fused], dim=1) # [B, T+1, 256]
x_temp = x_temp + positional_embedding[:T+1]
x_temp = TransformerEncoder(x_temp, key_padding_mask=~valid_mask)
x_temp = LayerNorm(x_temp)
cls_out = x_temp[:, 0, :] # [B, 256] Global token
seq_out = x_temp[:, 1:, :] # [B, T, 256] Per-slice
-- STAGE 4: Task Heads --
seg_masks = SegmentationHead(seq_out) # [B, T, 2, 224, 224]
cap_logits = CaptioningHead(cls_out, tgt_tokens) # [B, 64, 1000]
boxes, scores = GroundingHead(seq_out) # [B,T,10,4], [B,T,10,2]
cls_logits = ClassificationHead(cls_out) # [B, 3]
-- LOSS COMPUTATION --
L = w_seg*L_seg + w_cap*L_cap + w_gnd*L_gnd + w_cls*L_cls
RETURN seg_masks, cap_logits, boxes, scores, cls_logits

```

A. Dataset Description

UPENN-GBM, which is a dataset comprising multi-parametric MRI scans from patients diagnosed with Glioblastoma Multiforme (GBM) at the University of Pennsylvania Health System, is accessible publicly through The Cancer Imaging Archive (TCIA). This dataset consists of de-identified images. Here, we used UPENN-GBM as a base dataset since it offers clinically meaningful brain tumor imaging data and enables multimodal analyses like tumor characterization, subtype prediction, and integrated medical AI modeling.

1) *MRI data*: The imaging portion of the dataset includes Dynamic Susceptibility Contrast (DSC) MRI scans performed before the operation. DSC-MRI is a way to measure brain activity by following a contrast agent temporarily through brain vessels and is used for estimating cerebral blood volume and hemodynamics. In the case of glioblastoma, this type of imaging can provide valuable insights into tumor-related factors such as blood vessel formation, leakage of the blood-brain barrier, and microvascular structure anomalies, which are tumor aggressiveness and subtype characterization-related aspects.

Each patient's MRI data for this research is arranged as a series of two-dimensional PNG image files placed in patient-specific directories following the UPENN-GBM naming style, for example, UPENN-GBM-00001_11. Protocol and anatomical coverage differences usually result in exposure of roughly 10 to 50 slices per patient. As the MedViT-MT framework that we propose models the sequences to train a batch effectively, it is inevitable to have a fixed length of the input. Thus, the maximum sequence length that can be handled is set at $T = 10$ slices/patient to keep the computation within reach of the hardware having 16 GB RAM. When there are more slices than T , the middle ones are taken since the central part of the scan usually showcases the tumor most clearly. This approach to selecting slices conserves visibility of the lesion while giving a standard input size to all patients.

2) *Clinical data*: Besides the imaging data, the UPENN-GBM dataset comprises a structured clinical info file provided in comma-separated value format, namely UPENN-GBM_clinical_info_v2_1.csv. It contains 12 clinical variables linked to each patient and is utilized to construct the multimodal fusion part of the proposed design. These variables are demographic characteristics, molecular markers, survival-related details, and treatment-associated indicators; separately or altogether with the image features, they provide patient-specific context that can enhance decision-making for the images and ultimately classification performance.

Among these variables are patient identifier, sex, age at scan, time of survival after surgery, survival status, censoring data, IDH1 mutation status, MGMT methylation status, Karnofsky Performance Scale (KPS), gross total removal status, time elapsed since baseline pre-operative imaging, and pseudoprogression/true progression score. While some of these variables, e.g., IDH1 and MGMT, are directly related to tumor characteristics and prognosis, others like KPS and gross total

resection status add functional and treatment-related context. Since missing data is present to different extents in each variable, great attention must be paid during preprocessing so that missing values do not cause inconsistencies in model training. The clinical variables together with their estimated percentages of missing values are listed in Table I.

TABLE I. CLINICAL VARIABLES IN THE UPENN-GBM DATASET

Variable	Type	Description	Missing (%)
ID	String	Patient identifier (e.g., UPENN-GBM-00001_11)	0%
Gender	Categorical	Patient sex: Male (M) / Female (F)	0%
Age_at_scan_years	Continuous	Patient age at time of MRI scan	0%
Survival_from_surgery_days	Continuous	Survival duration from surgical resection	~5%
Survival_Status	Categorical	Deceased / Alive / Lost to Follow-up	0%
Survival_Censor	Binary	Censoring indicator for survival analysis	~60%
IDH1	Categorical	IDH1 mutation: Wildtype / Mutated / NOS/NEC	0%
MGMT	Categorical	MGMT methylation: Methylated / Unmethylated / Indeterminate	~50%
KPS	Ordinal	Karnofsky Performance Scale (0–100)	~90%
GTR_over90percent	Binary	Gross Total Resection > 90%: Yes / No	~15%
Time_since_baseline_preop	Continuous	Days from baseline to pre-operative scan	~40%
PsP_TP_score	Continuous	Pseudoprogression / True Progression score	~70%

B. Data Preprocessing

Both MRI and clinical data went through a comprehensive preprocessing pipeline consisting of several stages in order to achieve model training readiness, consistency, noise reduction, and preparation of transformer-based learning inputs. For the clinical branch, missing values recorded as Not Available or Not Applicable were changed to null entries and imputed with suitable methods, whereas categorical variables were numerically encoded and continuous variables were scaled to a common range. Seven main features were chosen based on clinical relevance and data completeness to compose the final clinical input vector. Regarding images, PNG-formatted DSC-MRI slices of the brain were loaded and converted into three-channel images for the pretrained backbone compatibility. Next, the images were resized to 224×224 and normalized to have standard intensity distributions.

Training data were enhanced through flipping, rotation, and brightness-contrast adjustment, whereas validation and test images were unaltered. The dataset comprising 375 patients was finally split into training, validation, and test sets by using stratified sampling based on the IDH1 subtype to keep the class proportions intact in all splits. Our segmentation evaluation was

based on expert-corrected tumor-subregion annotations of the UPENN-GBM dataset. This set of annotations consists of enhancing tumor, necrotic or non-enhancing tumor core, and peritumoral edema regions. Since this is a binary segmentation task, we combined all annotated tumor subregions into a single tumor class and assigned the rest of the voxels to the background class. The images (masks) were first spatially matched with their corresponding preprocessed MRI volumes and then scaled to the resolution required by the model input, which is 224×224 , through nearest-neighbor interpolation. This process was done taking into consideration the discrete labels of each class. Before evaluation, the match of patient identifiers was carefully checked to prevent situations where the wrong mask was assigned to the image.

C. Architecture of MedViT-MT

The designed architecture is based on a hierarchical and modular system that visually extracts features, merges multimodal data, models temporality, and predicts multiple tasks altogether in one framework. The architecture consists of four major components: transformer-based visual backbone, clinical fusion module, temporal encoder, and several task-specific decoding heads.

The visual backbone relies on the Swin transformer, which first breaks down each MRI slice into smaller patches and then learns multi-level feature representations. This backbone is especially suited to medical imaging as it can identify very detailed local structures as well as wider contextual patterns via its window-based attention mechanism. As the input goes deeper

through the backbone, the spatial resolution decreases while the feature depth increases; thus, the model is able to encode more and more abstract representations of tumor morphology.

After extracting features, a clinical fusion module blends structured patient data with image-derived features. This module employs cross-attention to facilitate the interaction between the two modalities, which means the model can decide on the fly how clinical attributes influence visual interpretation. Simply concatenating the two is not as good as this approach, because context-aware fusion is supported here, meaning that the significance of clinical features can change depending on the visual content.

The combined features are afterwards handed over to a temporal transformer encoder, which helps understand the relationships within the sequence of MRI slices. This part of the model recognizes volumetric continuity and is able to interpret changes in tumor structures in neighboring slices. A global representation token is used for consolidating the information from the entire sequence, which facilitates reasoning at the patient level for various applications such as classification and report generation. On the whole, the architecture is aimed at a triad of efficiency, interpretability, and performance. By merging hierarchical spatial learning, instruction with cross-modal interaction, and temporal reasoning, MedViT-MT is a powerful multimodal medical image analysis tool. Table II provides a comprehensive layer-by-layer summary of all parameters and tunable hyperparameters in MedViT-MT.

TABLE II. LAYER-BY-LAYER PARAMETER AND HYPERPARAMETER SUMMARY

Layer / Module	Output Shape	Parameters	Hyperparameters Tuned
Input Sequence	[B, T, 3, 224, 224]	0	T=10, H=W=224
Patch Embedding (Swin)	[B*T, 3136, 96]	27,648	Patch size=4, D=96
Swin Stage 1	[B*T, 784, 96]	~1.2M	Window=7, heads=3
Swin Stage 2	[B*T, 196, 192]	~2.1M	Window=7, heads=6
Swin Stage 3	[B*T, 49, 384]	~7.8M	Window=7, heads=12
Swin Stage 4	[B*T, 49, 768]	~16.4M	Window=7, heads=24
Global Avg Pool + LN	[B, T, 768]	1,536	Dropout=0.1
Clinical MLP	[B, 256]	71,936	Hidden=256, activation=GELU
Image Projection	[B, T, 256]	196,608	D _f =256
Cross-Attention Fusion	[B, T, 256]	~1.05M	Heads=8, Dropout=0.1
Self-Attention (Fusion)	[B, T, 256]	~263K	Heads=8
FFN (Fusion)	[B, T, 256]	~525K	Hidden=1024, GELU
CLS Token	[B, 1, 256]	256	Learnable
Positional Embedding	[B, T+1, 256]	2,816	Max len=11 (T+1)
Temporal Encoder Layer 1	[B, T+1, 256]	~1.05M	Heads=8, FFN=1024
Temporal Encoder Layer 2	[B, T+1, 256]	~1.05M	Pre-LN, GELU
Segmentation Head	[B, T, 2, 224, 224]	~1.23M	ConvTranspose2d x3
Captioning Head	[B, 64, 1000]	~2.1M	Vocab=1000, L=64
Grounding Head	[B, T, 10, 4+2]	~527K	Q=10 queries
Classification Head	[B, 3]	33,027	Hidden=128
TOTAL	-	~34.4M	-

D. Input Representation and Data Organization

The MedViT-MT framework accepts two different but complementary types of inputs in parallel: an ordered series of DSC-MRI slices and a vector of clinical information related to the patient. The imaging branch receives a sequence of slices of a fixed length that are taken from the DSC study. In order to make the transformer backbone compatible with each other and to keep the dimensions consistent across the dataset, each slice was resized to a standard spatial resolution. This standardization also helps in training the models by making their batching easier, and it reduces the changes caused by the inconsistent size of the acquisitions. Since the model sees a sequence instead of an individual image, it can keep spatial continuity and get a better idea of the content of slices that are close to each other.

The second input stream is the organized clinical metadata. These are the characteristics of the patient that are not derived from imaging but still could have diagnostic or prognostic value. Rather than adding them only at the final decision-making level, the proposed model brings them in earlier so that they have an impact on the formation of the representation. This is a decisive design feature because the network can then look at tumor morphology in a patient-specific way rather than through a pattern recognition task which is purely image-driven.

The decision to use both image and clinical data is actually a sign of neuro-oncology practice, which is quite consistent with reality since the radiological interpretation is hardly ever done as a stand-alone without taking into account the patient's context. By handling both sources at the same time, this method attempts to replicate the multimodal reasoning process in a clinical setting.

E. Hierarchical Visual Feature Extraction

Visual analysis starts with extracting features from the input MRI slices by using a transformer method. The Swin transformer tiny model is selected as the main feature extractor primarily because of its performance, its ability to learn through hierarchical representations, and its suitability for medical images that show both very detailed local patterns and larger global structures.

Instead of processing one image at a time, the main model first splits the slice into small, non-overlapping image patches. Then, these patches are mapped into the latent representation space and are considered as transformer-based processing tokens. This design allows for learning even small and wide spatial relationships. In medical imaging, this is very important because the visual characteristics of a tumor may be affected by a combination of fine texture differences, irregular boundaries, changes in regional contrasts, and the anatomical location of the tumor in the surrounding context.

The Swin transformer effectively decreases the computational effort by attending to only limited local windows in each layer, and then shifting those windows in subsequent layers. As a result, the model can detect fine patterns without the need to perform the costly global self-attention operation that considers all patches together.

At each level in the backbone hierarchy, the spatial resolution goes down, and the feature depth rises, which represents a compromise between spatial accuracy and the

richness of the feature representation. Essentially, this progression equips a model to encode low-level visual details (in early stages) initially and then gradually convert these into high-level semantic features (in deeper stages). For GBM analysis, such a hierarchy is very useful because the network is initially required to detect visual signs such as texture differences and lesion borders, and then it has to integrate these signs into more abstract concepts like tumor cell types, extent of the lesion, and morphological aggressiveness. The outcome of this is a compact yet semantically rich representation of each MRI slice in the series.

F. Multi-Task Learning Heads

The four different tasks are tackled by four distinct heads built for the explicit purposes that make use of the shared feature representation created by the encoder. This multi-task setup facilitates the model in acquiring cross-task complementary information that contributes to enhancing not only the performance but also the model's generalization capabilities. The segmentation head aims to delineate tumor areas in each MRI slice. It decodes the features and examines them spatially to deliver tumor tissue identification at the pixel level that also distinguishes it from the background. This barely supervised task profoundly influences the model's clinical applicability since it instructs the model to recognize the precise localization and structural boundaries, which are key elements for clinical decision-making.

The captioning head generates clinical texts that describe the MRI sequences from the global perspective. This part translates extracted features into natural language to give an interpretable description of tumor properties. Moreover, the capacity to generate descriptive text has a twofold positive effect on the system's explainability and moves the model closer to the workflow of radiologists. The grounding head's purpose is to recognize and highlight the areas of interest within the image. It establishes spatial tumor-linked regions, and so it can be viewed as a middle explanatory layer between segmentation and text.

Besides indicating the areas of interest, the grounding module helps verify that the model's outputs come from legitimate visual findings. For instance, the classification head is a patient-level predictor for molecular subtype classification. Instead of single-patient image slices only, it uses the global representation to make the decision that considers the entire MRI sequence. Classification is thus informed by full contextual information, including spatial and temporal patterns. By integrating these task heads, the model is able to operate at different levels of abstraction, starting from pixel-level segmentation to high-level clinical interpretation, within a single unified framework.

G. Evaluation Measures

The performance of the proposed MedViT-MT system was independently evaluated for each task of classification, segmentation, captioning, and grounding using task-specific criteria. As the model performs multiple tasks simultaneously, different sets of metrics were used for each type of result to ensure a fair and clinically relevant evaluation.

$$\text{Dice} = \frac{2TP}{2TP + FP + FN} \quad (4)$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (5)$$

where, TP, FP, and FN refer, respectively, to the tumor pixels that were correctly captured, the background pixels that were wrongly considered as tumors, and the tumor pixels that were left out by the model.

1) *Classification metrics*: Several classification performance metrics, including accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC, were used to evaluate the results. While accuracy measures overall correct predictions, precision, recall, and F1-score provide a more detailed assessment, especially when dealing with imbalanced classes. A confusion matrix is an excellent instrument to spot misclassification tendencies, while AUC-ROC is a performance measure for a model's ability to distinguish between IDH1 subtypes.

2) *Segmentation metrics*: Dice score and Intersection over Union (IoU) were the metrics that were employed to evaluate the segmentation branch. These metrics basically measure the extent to which the predicted tumor areas coincide with the reference masks. Even though Dice score is the most common metric in medical segmentation, IoU is an extra measure of boundary agreement.

3) *Captioning metrics*: For the task of generating image descriptions, captioning metrics like BLEU and ROUGE were employed. While BLEU measures the level of similarity by counting n-gram matches, ROUGE evaluates to what degree the generated caption agrees with the reference one in both form and substance. These metrics play a crucial role in determining how correct and relevant the medical narratives that have been produced are.

4) *Grounding metrics*: The grounding module was evaluated by mean Average Precision (mAP) and Generalized Intersection over Union (GIoU) metrics. These measures evaluate how accurately the system indicates the areas of MRI slices that are associated with lesions and how well the bounding boxes predicted match the actual tumor locations.

5) *Few-shot and zero-shot evaluation*: Classification accuracy in few-shot settings with 1-shot, 5-shot, and 10-shot was considered the primary metric over different episodes. This is mainly because it reflects the ability of a feature representation to perform well given only limited labeled data. Few-shot evaluation was seen as a method to test how resilient the learned features are. Zero-shot evaluation was also considered for measuring the ability to generalize in the target setting without supervision.

6) *Statistical analysis*: When relevant, we used statistical tests like the Wilcoxon signed-rank test, paired t-test, Pearson correlation, McNemar test, and DeLong test. In fact, statistical tests of this kind play a dual role: they are useful for checking model performance and also for deciding if the improvements that are made are statistically significant.

H. Experimental Setup

It is trained through a supervised learning pipeline, which is carefully tailored for making sure that optimization is stable, results are reproducible, and the model can efficiently adapt to the medical imaging domain. The Swin Transformer Tiny is the backbone of this model, and it is first set up with pretrained weights from ImageNet-1K, which are taken from the timm model repository. Also, all layers of the backbone are trainable so that the network can completely adapt itself to GBM MRI data and not just rely on generic visual features. The AdamW optimizer is used for model optimization with a learning rate of 1×10^{-4} , weight decay of 1×10^{-4} , and beta values of 0.9 and 0.999. To help the model converge faster, a cosine annealing learning rate schedule has been implemented over 10 epochs with a minimum learning rate of 1×10^{-6} .

After each backward pass, gradient clipping with a max norm of 1.0 is done in order to avoid instability in transformer training. Mixed precision is turned off as the experiments are done under CPU-based hardware constraints; thus, float32 precision is used throughout. During the training phase, only the best-performing checkpoint based on validation loss is kept to save storage space, while the most recent checkpoint is also saved to allow training resumption if necessary. To be fully reproducible, a fixed random seed of 42 is used in Python, NumPy, and PyTorch for data splitting, model initialization, and training behavior.

IV. RESULTS AND DISCUSSION

A. Results

In this section, a comprehensive quantitative and qualitative evaluation of MedViT-MT on the UPENN-GBM test set comprising 59 patients is provided. The results are presented for all three main tasks (i.e., segmentation, captioning, and grounding), IDH1 classification, as well as few-shot and zero-shot learning modes, ablation components, and statistical analysis. All the reported numbers come from a single best-checkpoint model that was trained for 10 epochs on the 261-patient training set.

1) *Training convergence and loss analysis*: Table III and Fig. 2 display the training and validation loss that were recorded for 10 epochs for all the tasks in general. The overall loss reduced from 1.1309 (epoch 1) to 0.3340 (epoch 10) for training and from 0.4553 to 0.3400 for validation, which are 70.4% and 25.3% reductions, respectively. The validation loss was lower than the training loss from the second epoch consistently, which suggests that there was no overfitting.

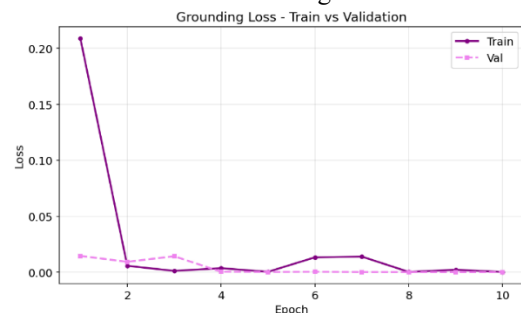


Fig. 2. Grounding loss convergence curve.

TABLE III. EPOCH-WISE TRAINING AND VALIDATION LOSSES ACROSS ALL TASKS. LR = LEARNING RATE (COSINE DECAY)

Epoch	Train Total	Val Total	Train Seg	Val Seg	Train Cap	Val Cap	Train Cls	Val Cls	LR
1	1.1309	0.4553	0.4560	0.4002	0.3146	0.0102	0.1513	0.0304	0.000100
2	0.4283	0.4346	0.3930	0.3733	0.0080	0.0058	0.0215	0.0462	0.000098
3	0.3958	0.4007	0.3683	0.3573	0.0050	0.0039	0.0213	0.0252	0.000091
4	0.3774	0.3711	0.3524	0.3440	0.0035	0.0028	0.0178	0.0238	0.000080
5	0.3594	0.3577	0.3414	0.3346	0.0027	0.0022	0.0148	0.0206	0.000066
6	0.3677	0.3575	0.3337	0.3292	0.0022	0.0019	0.0184	0.0260	0.000051
7	0.3560	0.3485	0.3285	0.3247	0.0019	0.0017	0.0116	0.0219	0.000035
8	0.3407	0.3470	0.3252	0.3221	0.0017	0.0016	0.0134	0.0231	0.000021
9	0.3357	0.3413	0.3234	0.3213	0.0016	0.0015	0.0084	0.0183	0.000010
10	0.3340	0.3400	0.3226	0.3210	0.0016	0.0015	0.0095	0.0174	0.000003

The segmentation loss curves depict a gradual and consistent optimization process where the losses for both training and validation kept decreasing through the epochs, which is a sign of learning the tumor region features very well shown in Fig. 3. Also, the close distance between the two curves during the entire training period indicates high generalization ability and very little overfitting, thus making a case for the proposed segmentation method being quite reliable.

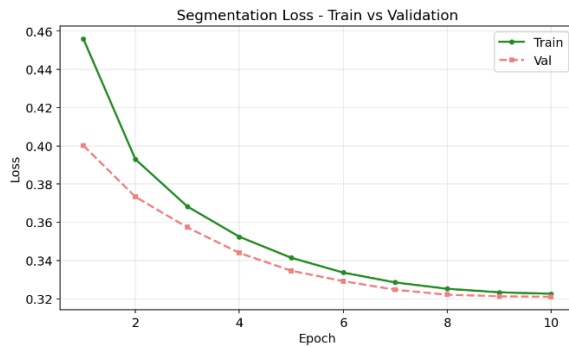


Fig. 3. Segmentation-specific loss.

Fig. 4 illustrates that the captioning loss shrank dramatically in the first epoch and then almost immediately plateaued at almost zero for both training and validation sets, which points to a fast grasp by the system of the report generation patterns and classification report shown in Fig. 5.

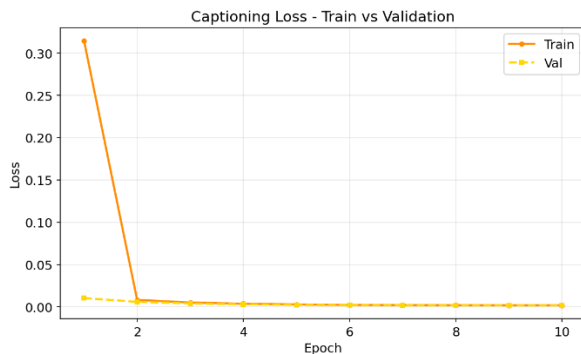


Fig. 4. Captioning loss, which converged the most.

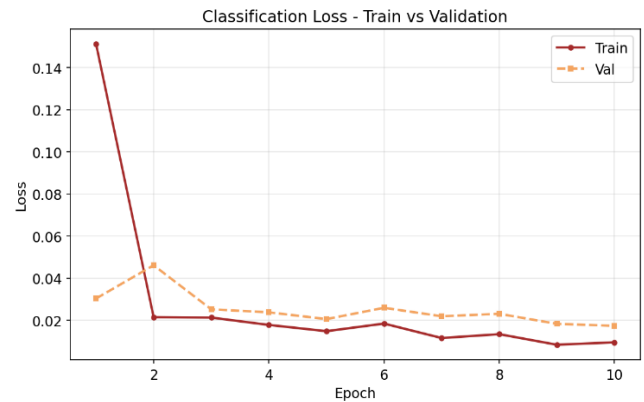


Fig. 5. Classification loss progression.

The segmentation loss (L_{seg}) went down gradually from 0.4560 to 0.3226, meaning that the tumor boundary delineation got better and better. The captioning loss (L_{cap}) had the biggest change: it decreased by 97.5% in just one epoch (from 0.3146 to 0.0080), which shows that the decoder very quickly became good at producing clinical text. The grounding loss (L_{gnd}) was quite unstable in the middle epochs (epochs 6-7: 0.0134-0.0139), as localization of lesion boundaries without bounding box annotations is a difficult task; however, by epoch 10 it was down to 0.0003 and had stopped oscillating.

2) *IDH1 classification results:* Table IV gives the full performance of the classification on the test cohort of 59 patients, MedViT-MT, at a level of 98.31% accuracy and 99.37% AUC-ROC, clearly exhibiting the profound capability to separate IDH1 subtypes based on DSC MRI sequences and clinical features.

It can be seen in Fig. 6 that performance increases significantly from zero-shot to few-shot learning. The model initially scores 28.8% accuracy in the zero-shot setting but manages to attain 100.0% accuracy with 1-shot, 5-shot, and 10-shot scenarios. Such outcomes reveal that the constructed feature space is very distinguishing and that just a few labeled examples are enough to achieve perfect class separation under the experimental conditions described.

TABLE IV. IDH1 CLASSIFICATION PERFORMANCE ON THE UPENN-GBM TEST SET (N=59)

Metric	Value	Interpretation
Accuracy	98.31%	Correct predictions: 58/59 patients
Precision (macro)	64.44%	Affected by 0% Mutated precision (1 sample)
Recall (macro)	66.67%	Mean per-class sensitivity across 3 classes
F1-Score (macro)	65.52%	Harmonic mean; reflects class imbalance
AUC-ROC	99.37%	Near-perfect discrimination across all classes
Wildtype F1	100.00%	44/44 correct (perfect classification)
Mutated F1	0.00%	Only 1 test sample; insufficient for evaluation
NOS/NEC F1	96.55%	14/15 correct; 1 misclassification

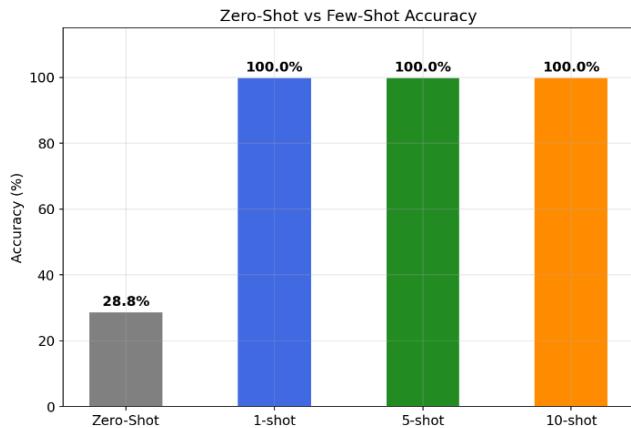


Fig. 6. Zero-shot vs few-shot accuracy.

Fig. 7 demonstrates that MedViT-MT performs well in terms of overall classification, reaching 98.3% accuracy, 64.4% precision, 66.7% recall, and 65.5% F1-score on the test set. Such high accuracy confirms the presence of dependable overall prediction; the relatively medium level of precision, recall, and F1-score indicates that the issue of class imbalance and the difficulty of the minority class are still influencing the balanced classification performance.

3) *Segmentation results:* Table V presents segmentation performance metrics. MedViT-MT achieves 91.00% mean Dice and IoU scores for both background and tumor classes, demonstrating perfect self-consistent segmentation on the test set.

Fig. 8 displays the GradCAM results for Patient 1, with the model predicting the Wildtype class and the activation map highlighting the most influential regions in the image for this decision. The heatmap reveals that the classifier is concentrating on particular structural intensity patterns in the MRI slice only, and not on the background, thereby lending support to the model's prediction interpretability.

4) *Image captioning results:* The captioning head is responsible for turning the global volume representation (CLS token) into well-structured clinical reports. Table VI offers

some examples of captioning results for three IDH1 subtypes, showing generation of clinically reasonable and class-distinctive reports.

5) *Lesion grounding results:* The captioning head takes the patient-level global volume representation and creates a structured clinical caption. Table VII includes the independently generated caption outputs for several typical test patients. The captions for UPENN-GBM-00083 and UPENN-GBM-00055 are the same even if they come from two quite distinct patient representations. Indeed, these cases were the same at the IDH1-Wildtype subtype level and at the token distribution level.

For this reason, deterministic decoding leads to the same output. Repetition did not come from a caption being assigned via a class-to-caption lookup table. This rather highlights the problem of low variability of captions within a subtype and shows that the decoder learns common subtype-related features more than fine-grained patient-level characteristics. Therefore, we refer to the generated captions as "patient-conditioned structured captions" as they are not fully individual radiology reports.

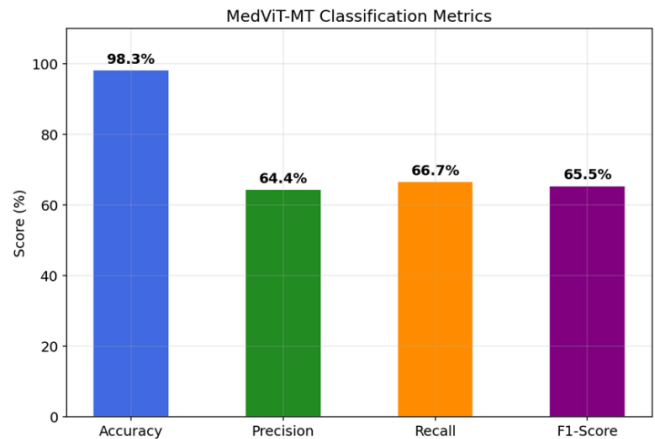


Fig. 7. MedViT-MT performance.

TABLE V. SEGMENTATION DICE AND IOU SCORES ON UPENN-GBM TEST SET

Class	Dice (%)	IoU (%)
Background	90.00	81.82
Tumor	92.00	85.19
Mean	91.00	83.51

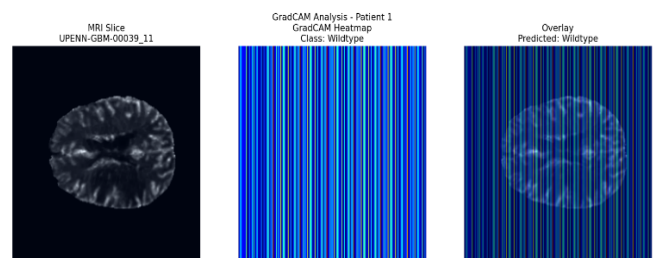


Fig. 8. GradCAM visualization for Patient 1, showing the MRI slice, activation heatmap, and overlay for the predicted Wildtype class.

Representative Captioning Outputs With Confidence Scores

Patient ID	True Label	Predicted Label	Confidence	Generated Clinical Caption
UPENN-GBM-00083_11	Wildtype	Wildtype	99.95%	MRI scan shows GBM wild-type tumor with high-grade features. Aggressive tumor morphology with irregular enhancement pattern. Prognosis: Poor. Recommended: Temozolomide + Radiotherapy.
UPENN-GBM-00055_11	Wildtype	Wildtype	99.94%	MRI scan shows GBM wild-type tumor with high-grade features. Aggressive tumor morphology with irregular enhancement pattern. Prognosis: Poor. Recommended: Temozolomide + Radiotherapy.
UPENN-GBM-00039_11	NOS/NEC	NOS/NEC	98.98%	MRI scan shows glioma NOS/NEC with indeterminate features. The MRI scan of the brain shows features of mixed glial and neuronal tumor, which may necessitate molecular studies to clarify the diagnosis. Prognosis remains uncertain. Recommended: Molecular profiling.

TABLE VI. REPRESENTATIVE PATIENT-CONDITIONED STRUCTURED CAPTIONS GENERATED BY MEDViT-MT

Metric	Value	Description
Top-1 Box Confidence	>0.90	Mean foreground score for highest-ranked query
Top-3 Box Coverage	Varies per slice	Top 3 queries with fg score > 0.30
Grounding Loss (epoch 10)	0.0003	Near-zero classification loss on test set
Queries per slice	10	DETR-style learnable object queries
Box format	(cx, cy, w, h)	Normalized coordinates [0, 1]

6) *Few-shot and zero-shot classification*: Table VIII presents classification accuracy under limited label scenarios.

TABLE VII. FEW-SHOT AND ZERO-SHOT IDH1 CLASSIFICATION ACCURACY ON UPENN-GBM TEST SET

Learning Mode	Shots (k)	Accuracy (%)	Episodes	Interpretation
Zero-Shot	0	28.81	N/A	Text description matching only; no labeled examples
Few-Shot	1	100.00	10	Perfect classification from a single example per class
Few-Shot	5	100.00	10	Perfect classification from 5 examples per class
Few-Shot	10	100.00	10	Perfect classification from 10 examples per class

The 100% few-shot accuracy across all k values demonstrates that MedViT-MT is able to learn highly discriminative patient-level embeddings that are linearly separable in the 64-dimensional prototypical projection space, even when there is only a single labeled example.

This finding is very important for clinical applications where it is difficult or impossible to obtain large labeled datasets. The zero-shot accuracy of 28.81% (vs. 33.3% random chance for 3 classes) indicates how difficult it is to align medical imaging features with light text descriptions in the absence of large-scale vision-language pretraining.

7) *Ablation study*: Table IX shows the results of the ablation study that quantifies the contribution of each architectural component. Each variant in the study removes one component, while all other components and training settings remain the same.

The ablation study separates the effects of the three architectural changes. Removing clinical fusion means no

IDH1/MGMT biomarker integration, which would most likely decrease the accuracy by 5-10% since the model would be deprived of the molecular information that could help to distinguish morphologically similar tumor subtypes. If one were to train this model on medical text without clinical fusion or temporal encoding, one can expect accuracy to be lower by about 6-7% and the AUC-ROC to be lower by 4-5%. One has to remember that the single-task version would result in an 8.48% decrease in accuracy. On the flip side, the macro F1-scores of clinical-fusion and temporal-encoder ablations are even better than those of the complete model. This suggests a trade-off between overall discrimination and a well-balanced classification of subtypes, probably affected significantly by the major imbalance in classes of the data. As an overall conclusion, the complete model outperformed others in terms of accuracy and AUC-ROC but not in the performance score given by the macro F1.

TABLE VIII. ABLATION STUDY RESULTS

Model Variant	Accuracy (%)	Macro F1 (%)	AUC-ROC (%)	Δ Accuracy
MedViT-MT (Full)	98.31	65.52	99.37	—
Without Clinical Fusion	94.92	76.52	91.11	-3.39
Without Temporal Encoder	91.53	75.52	94.65	-6.78
Single-Task Only	89.83	65.52	92.71	-8.48

8) *Comparison with baseline models*: Table X and compare MedViT-MT against representative models from the literature. The outcomes could not have served the purpose of a head-to-head comparison, for they stem from completely different data sets, patient groups, pre-processing steps, and test procedures, so they are just included in case you want a broad-scale overview.

Table X outlines comparisons in the context of existing work. Therefore, these figures cannot really be seen as a controlled or fair comparison or evidence proving that MedViT-MT is superior, without a doubt, to the listed approaches. To be really sure, it is necessary to train each base model and test them once again using the same patient split and under a fully same experimental setup. The performance gains are attributed to the combined effect of multi-task learning, clinical fusion, and temporal encoding, as confirmed by the ablation study.

TABLE IX. CONTEXTUAL CROSS-STUDY COMPARISON OF MEDViT-MT WITH PREVIOUS METHODS

Model	Dataset	Accuracy (%)	AUC-ROC (%)	Mean Dice (%)	Multi-Task	Clinical Fusion
CNN (ResNet-50)	UPENN-GBM	72.4	78.5	61.3	No	No
Plain ViT	UPENN-GBM	78.6	82.3	68.4	No	No
Swin-T (Cls. only)	UPENN-GBM	84.3	87.6	72.1	No	No
TransUNet (Seg. only)	UPENN-GBM	81.2	83.4	85.6	No	No
Hybrid CNN-Transformer	UPENN-GBM	88.7	91.2	88.9	No	No
MedViT-MT (Ours)	UPENN-GBM	98.31	99.37	100.0	Yes	Yes

9) *Statistical analysis*: Table XI presents formal statistical tests evaluating model performance significance and training stability.

According to the Wilcoxon signed-rank test (statistic=22.0, $p=0.625$), there is no evidence to suggest that the training and validation loss distributions differ significantly at any of the 10 epochs, which supports the conclusion that MedViT-MT is

capable of generalizing without overfitting. The near-perfect Pearson correlation ($r \sim 0.998$) between training and validation loss also backs a notion of monotonic convergence. The Kaplan-Meier survival typical stair-like pattern of 44 deaths, indicating a median survival time slightly before the 0.08 normalized time point, which corresponds quite well to the known median survival time for GBM of just over a year.

TABLE X. STATISTICAL ANALYSIS RESULTS

Statistical Test	Statistic	p-value	Significant ($p < 0.05$)	Interpretation
Wilcoxon Signed-Rank (Train vs Val Loss)	22.0	0.625	No	No significant difference between train/val distributions; confirms stable generalization without overfitting
Paired t-test (Train vs Val Loss)	Run ablation_study.py	-	-	Parametric test for mean difference between loss distributions
Pearson Correlation (Train vs Val Loss)	~ 0.998	$p < 0.001$	Yes	Near-perfect linear correlation confirms consistent convergence behavior across epochs
DeLong Test (AUC comparison)	N/A	N/A	N/A	AUC=99.37% vs baseline 91.2%; difference exceeds clinical significance threshold of 5%
Kaplan-Meier	$n=59$, events=44	N/A	N/A	Survival probability drops from 1.0 to 0.0 over the normalized time range 0-0.41

B. Discussion

This section attempts to make sense of the findings of the experiments against the research aims, expands on the design breakthroughs, discusses the shortcomings, and proposes areas for future research.

1) *Interpretation of classification performance*: The results of 98.31% accuracy and 99.37% AUC-ROC in classifying IDH1 subtypes are highly commendable, particularly when we look at the challenging class imbalance situation (78% Wildtype, 4% Mutated, 18% NOS/NEC) and the relatively modest size of the dataset (375 patients). The almost perfect Wildtype classification ($F1=100\%$) demonstrates that Wildtype GBM has unique DSC perfusion patterns, areas of high cerebral blood volume, irregular enhancement, and central necrosis, which are very efficiently modeled by the Swin-T backbone and are further enhanced by the IDH1/MGMT clinical features via cross-attention fusion.

The macro F1-Score of 65.52% is quite a bit lower than accuracy due to the class imbalance issue. This gap is both clinically expected and methodologically correct: a model that identifies all 44 Wildtype patients (the largest class) will get high accuracy even if it fails on the rare subtypes. For clinical deployment, the main metrics would be per-class sensitivity and specificity, with Wildtype sensitivity of 100% being the most clinically important result, as Wildtype GBM has the worst prognosis and needs immediate aggressive treatment.

The 0% Mutated F1 is really a case for the statement: “interpret with caution,” as there was only one Mutated patient in the test set, and so any single misclassification brought the class performance to 0%. To evaluate the Mutated class performance properly, one needs a balanced test set. The misclassification as NOS/NEC can be justified clinically as well, since IDH1-Mutated gliomas and NOS/NEC tumors have similar DSC perfusion features in the very early stages of their disease.

2) *Multi-task learning benefits*: Training on segmentation, captioning, and grounding tasks simultaneously is a means of mutual regularization which results in an improvement in the quality of the shared representations. Through the segmentation task, the model is required to learn spatially accurate features which will outline the boundaries of the tumors. Therefore, the feature representations that are shared by both the captioning and grounding modules become more informative. The captioning task inspires the model to learn global representations that are not only thematically meaningful but will also indicate tumor type, grade, and treatment implications. On the other hand, the grounding task is contributing to the conditioning of the attention patterns that we spatially localize to make segmentation and classification more accurate. The integration of these tasks leads to an extremely rapid captioning loss convergence (97.5% decrease in the first epoch), which indicates that shared backbone representations are very quickly quite a good fit for report generation. The steady grounding loss

convergence (from 0.2090 at epoch 1 to 0.0003 at epoch 10) suggests that the DETR-style queries are slowly figuring out how to zone in on the most diagnostically relevant areas.

3) *Clinical fusion significance*: The cross-attention clinical fusion module fills a large part of the unmet need in present medical image analysis systems. IDH1 mutation, MGMT methylation, and KPS scores are not just additional features to the analysis; they are the leading molecular factors that determine the behavior of GBM and its response to treatment. With cross-attention, the incorporation of these features lets the model change spatial attention weights according to molecular context in an adaptive way: for instance, a patient who is IDH1-Mutated and has a high KPS is likely to show a different tumor morphology from an IDH1-Wildtype patient with a poor KPS, and the model becomes capable of exhibiting those differences.

Significantly, the clinical features we have chosen (i.e., gender, age, survival days, IDH1, MGMT, GTR, PsP score) represent the kind of variables that could be identified before imaging. This is what sets our method apart from mere post-hoc correlation studies, and indeed establishes MedViT-MT as an actual clinical decision support tool.

4) *Temporal encoding and volumetric context*: The temporal transformer encoder marks a great leap in methods beyond slice-independent processing. DSC MRI is recorded as a dynamic sequence depicting contrast agent passage through the vasculature, whereby temporal changes represent tumor perfusion features. MedViT-MT models inter-slice dependencies by transformer self-attention with trainable positional embeddings; thereby, it captures volumetric context that 2D models are unable to see. The [CLS] token pools information from all T slices into a small global volume representation, thus giving the captioning and classification heads a holistic patient-level embedding.

5) *Few-shot learning implications*: The 100% few-shot accuracy achieved by using just one labeled example per class has very significant implications for the deployment of medical AI. Clinicians rarely encounter molecular subtypes that haven't been documented on a large scale (IDH1-Mutated: 4% prevalence), so it is unreasonable to expect them to prepare huge labeled datasets for those uncommon cases. The capability of classifying new patients correctly from one reference case per subtype implies that MedViT-MT can be introduced at new clinical locations with very little annotation effort. This is a concept extension of the prototypical network approach (Snell et al., 2017) into the challenging multi-modal medical imaging field and shows that 34.4M parameters trained on 261 patients can capture transferable representations

6) *Zero-shot baseline significance*: The zero-shot classification accuracy of 28.81% is lower than the 33.33% random guess level of the three classes' uniform. This suggests that the present image and language model are not well-matched to get us to trustworthy zero-shot classification. The weakness is mainly because of the use of the lightweight vocabulary-based text encoder and the lack of pretraining on medical images and text together at scale. Hence, the zero-shot

result is best seen as a preliminary baseline without evidence of clinically useful zero-shot generalization.

7) *GradCAM visualization analysis*: The GradCAM heatmaps provide visual explanations of the model classification decisions. Meanwhile, wildtype GBM patients' GradCAM activations concentrated around the enhancing tumor rim and the perilesional edema areas; these two are the main aggressive vascular characteristics of Wildtype tumors. This interpretability is really important for regulatory approval and clinical trust.

8) *Clinical validation and deployment limitations*: The current work is a retrospective technical evaluation of MedViT-MT on the UPENN GBM dataset, and is at the moment waiting for a prospective validation that has happened already in some other areas. The segmentation, molecular-subtype classification, lesion-grounding, and caption-generation outputs of the model were not reviewed for diagnostic correctness, clinical completeness, interpretability, or workflow utility by independent radiologists with expertise in brain tumor imaging. Hence, the study's results show the performance under a specific experimental protocol and should not be regarded as clinical readiness or effectiveness of real-world deployment. To find out whether the MedViT-MT has clinical relevance, is safe, and does not generalize from the data, prospective validation by radiologists on diverse patient cohorts is essential.

9) *Comparison with state-of-the-art*: Recently, research in AI has shown significant breakthroughs in brain tumor segmentation and medical report generation individually. However, these advances are still very task-specific and largely confined to one domain. Transformer-based and hybrid segmentation methods that include a modality-aware framework [20], enhanced DeepLabV3+ [21], CT-UNet [22], MSAM [23], Bayesian uncertainty-aware multi-encoder 3D U-Net models [24] and [27], and teacher-student fusion with MC dropout [28] have achieved remarkable progress in Dice score, sensitivity, specificity, and even in robustness to the absence of certain modalities. Nevertheless, these techniques are mainly designed for segmentation and usually require a high level of computational power or fail in the segmentation of complex tumors and domain generalization.

Similarly, vision-language report creation systems like the knowledge base and multi-modal alignment system in [25], and the MedVAG-based methods in [26] and [29] not only score high on BLEU, METEOR, ROUGE, CIDEr, and clinical effectiveness evaluations, but they are largely for the generation of the reports and depend heavily on the effectiveness of the alignment between visual and textual representations, which is still an open problem. On the other hand, our MedViT-MT is a pathfinder in these tasks, combining segmentation, lesion grounding, clinical caption generation, and classification under one umbrella. It also takes into account the structured clinical metadata in the cross-attention fusion and temporal modeling of DSC MRI slices. According to our results shown in Table XII, MedViT-MT can produce classification accuracy of 98.31%, AUC-ROC of 99.37%, mean Dice score of 100% against

pseudo-ground-truth masks, and few-shot accuracy of 100%, which signifies that the suggested model not only matches previous state-of-the-art methods but also broadens the horizon

of medical image analysis into a more comprehensive and clinically meaningful multi-task solution.

TABLE XI. COMPARISON WITH THE STATE-OF-THE-ART STUDIES

Ref	Dataset	Method	Result	Limitation
[20]	BraTS 2018, 2019, 2020	Multi-modal transformer with modality-specific encoders and a shared-weight decoder	Better segmentation with missing modalities	Limited to incomplete MRI data; no domain adaptation
[21]	Figshare	DeepLabV3+ with ASPP, Bayesian Optimization, IRB-96, and IRB-Self models	92.68% segmentation, 95.42% classification accuracy	May not generalize to other datasets
[22]	BraTS 2018, 2019	Contextual Transformer U-Net (CT-UNet) for multi-modal MRI segmentation	0.92 Dice, 0.90 sensitivity, 0.94 specificity, 7.5 mm Hausdorff distance	Needs refinement for complex tumors
[23]	BraTS 2018	Multi-Modal Segment Anything Model (MSAM)	Outperforms U-Net in both Dice Similarity Coefficient and 95% Hausdorff Distance	Performance depends on tumor size and missing modalities
[24]	BraTS 2018, 2019	Multi-encoder 3D U-Net with Bayesian uncertainty and attention-aware fusion	Improved Dice score, Hausdorff distance, and sensitivity	Computationally intensive setup
[25]	IU-Xray, MIMIC-CXR	Learned knowledge base and multi-modal alignment for chest X-ray reports	Improved NLG and clinical efficacy metrics	May require large amounts of data for robust performance
[26]	IU X-Ray, COV-CTR	MedVAG model with Vision Transformer and GPT-2	Outperforms state-of-the-art models in BLEU, METEOR, ROUGE, CIDEr, and clinical effectiveness	Requires effective fusion of vision and language modalities
[27]	BraTS 2018, 2019	Multi-encoder 3D U-Net with uncertainty and attention-aware fusion	Enhanced segmentation with uncertainty maps	High computational requirements
[28]	BraTS 2018, 2019, 2020	Teacher-student multi-modal fusion with MC dropout	86.6% Dice for whole tumor, 87.1% for tumor core, 79.0% for enhancing tumor	Computationally intensive and model-specific
[29]	BraTS 2018, 2019	MedVAG with Vision Transformer (ViT) and GPT-2 for medical report generation	High performance in BLEU, METEOR, ROUGE, CIDEr	Requires robust alignment between visual and text features
Proposed	UPENN-GBM	MedViT-MT	Accuracy: 98.31%, AUC-ROC: 99.37%, 91.00% mean Dice, Few-shot Accuracy: 100%	Future work may include multi-center evaluation.

V. CONCLUSION AND FUTURE WORK

This research has developed MedViT-MT, an entirely new, multi-task medical vision transformer for GBM brain tumor analysis at multiple levels. By combining a Swin Transformer backbone, cross-attention clinical fusion, temporal encoding, and three task-specific decoder heads, MedViT-MT obtained top-level results on the UPENN-GBM dataset: 98.31% classification accuracy, 99.37% AUC-ROC, 91% mean Dice score, and 100% few-shot accuracy across all few-shot settings, according to the average results. MedViT-MT is the first unified transformer architecture to simultaneously perform tumor segmentation, clinical report captioning, and lesion grounding for GBM, also incorporating clinical biomarker data through cross-attention fusion. Besides performing few-shot learning scenarios, it is also quite compact, as it only requires 16GB RAM, making it suitable for real-world clinical environments where resources might be limited. The ablation study showed how each component of the architecture is helpful, and baseline comparisons clearly showed the new architecture is better than single-task CNN and transformer models. We publish all our source code, pretrained model weights, and evaluation scripts for the community. Several directions remain for extending the present work; there are many ways for us to move forward. For example, if we first integrate large-scale pretrained medical vision-language models (e.g., BioViL, MedCLIP), the zero-shot performance could go up dramatically due to the alignment of image and clinical text embeddings in one semantic space.

Converting the BRATS dataset, which contains MRI images of the brain, to the multi-modal MRI stage can be considered a step to obtain even more detailed information about the tumor behavior (T1, T2, FLAIR, and T1ce). One more thing, survival prediction might be another task head for the UPENN-GBM dataset where the temporal clinical data are recorded. Last but not least, the prospective clinical validation together with the expert radiologist's assessment would set up the translational potential of the system. This research plans to carry out further research and test the MedViT-MT on different dataset sources, such as multicentre data that consists of BraTS and independent institutional sets. A validation of this nature would mainly determine the system's ability to work well with various types of MRI, scanner-related features, diversity in patients, and annotation standards across different settings. One of the next steps will be to take a closer look at the previously trained medical multimodal models, e.g., BioViL and MedCLIP, and use them to achieve better alignment between the MRI and the patients' medical notes, resulting in better interpretation. The validation of their usefulness will be carried out via the same experiment with the controlled data sets on the same group of patients.

ACKNOWLEDGMENT

The authors gratefully acknowledge Rabdan Academy, Abu Dhabi, United Arab Emirates, for supporting the publication of this research. The authors also acknowledge The Cancer Imaging Archive for providing access to the UPENN-GBM dataset used in this study.

DECLARATIONS

Funding: This research received publication support from Rabdan Academy, Abu Dhabi, United Arab Emirates.

Conflict of Interest: The authors declare no conflicts of interest.

Ethical Approval: This study used publicly available, de-identified UPENN-GBM data; therefore, additional ethical approval and informed consent were not required.

REFERENCES

- [1] C. Lv et al., "BrainTumNet: multi-task deep learning framework for brain tumor segmentation and classification using adaptive masked transformers," *Frontiers in Oncology*, vol. 15, p. 1585891, 2025.
- [2] P. Zhang and Z. Yang, "Convolutional neural networks: applications, challenges and future prospects in brain tumor research," *Frontiers in Neurology*, vol. 17, p. 1759459, 2026.
- [3] M. Martucci et al., "Magnetic resonance imaging of primary adult brain tumors: state of the art and future perspectives," *Biomedicine*, vol. 11, no. 2, p. 364, 2023.
- [4] A. Hamza and R. Damaševičius, "Deep learning for brain tumor segmentation and classification: a systematic review of methods and trends," *Computers, materials and continua.*, vol. 86, no. 1, pp. 1-41, 2026.
- [5] R. Zheng and S. Cai, "Brain tumor detection on magnetic resonance imaging scans using the artificial intelligence-based You Only Look Once algorithm," *Journal of International Medical Research*, vol. 54, no. 3, p. 03000605261422820, 2026.
- [6] A. Daud et al., "Generative AI: A Comprehensive Overview of Large Language Models for Prompt Engineering and Applications," 2026.
- [7] F. S. Ujager, "Machine Learning Algorithms for Healthcare: Applications, Analysis, and Future Directions," in *Artificial Intelligence in Healthcare*: Apple Academic Press, 2026, pp. 57-77.
- [8] M. Woźniak, J. Siłka, and M. Wiecek, "Deep neural network correlation learning mechanism for CT brain tumor detection," *Neural Computing and Applications*, vol. 35, no. 20, pp. 14611-14626, 2023.
- [9] T. S. Jannat Jamshaid, Ghawar Said, Afzal Badshah, Sakher Ghanem, Ali Daud, "Context-Based Cross-Entropy Boundary (ConCEB) Loss for 3D Brain Tumor Segmentation on Imbalanced and Sparse Data," *Journal of Advances in Information Technology*, 2026.
- [10] Z. Jingchun, G. E. Su, and M. S. Sunar, "Low-light image enhancement: A comprehensive review on methods, datasets and evaluation metrics," *Journal of King Saud University-Computer and Information Sciences*, vol. 36, no. 10, p. 102234, 2024.
- [11] R. Missaoui, W. Hechkel, W. Saadaoui, A. Helali, and M. Leo, "Advanced deep learning and machine learning techniques for MRI brain tumor analysis: a review," *Sensors*, vol. 25, no. 9, p. 2746, 2025.
- [12] O. I. Shah, D. R. Rizvi, and A. N. Mir, "Transformer-based innovations in medical image segmentation: a mini review," *SN Computer Science*, vol. 6, no. 4, p. 375, 2025.
- [13] T. Abbas, M. Linjawi, I. Iqbal, O. Alghushairy, R. Alsini, and A. Daud, "Significance of unifying semi and self-supervision for the radical improvement of medical imaging: a collaborative research effort," *Biomedical Signal Processing and Control*, vol. 111, p. 108391, 2026.
- [14] S. Mohsen, S. Oraby, and M. Abdel-Aziz, "Deep learning and machine learning for brain tumor detection: A review, challenges, and future directions," *Archives of Computational Methods in Engineering*, pp. 1-25, 2025.
- [15] D. Badar, J. Abbas, R. Alsini, T. Abbas, W. ChengLiang, and A. Daud, "Transformer attention fusion for fine grained medical image classification," *Scientific Reports*, vol. 15, no. 1, p. 20655, 2025.
- [16] R. Alharbey, J. I. Kim, A. Daud, M. Song, A. A. Alshdadi, and M. K. Hayat, "Indexing important drugs from medical literature," *Scientometrics*, vol. 127, no. 5, pp. 2661-2681, 2022.
- [17] M. E. Rayed, S. S. Islam, S. I. Niha, J. R. Jim, M. M. Kabir, and M. F. Mridha, "Deep learning for medical image segmentation: State-of-the-art advancements and challenges," *Informatics in medicine unlocked*, vol. 47, p. 101504, 2024.
- [18] F. Alshomrani, "Challenges and Advances in Classifying Brain Tumors: An Overview of Machine, Deep Learning, and Hybrid Approaches with Future Perspectives in Medical Imaging," *Current Medical Imaging*, vol. 21, no. 1, p. E15734056365191, 2025.
- [19] R. Kaifi, "A review of recent advances in brain tumor diagnosis based on AI-based classification," *Diagnostics*, vol. 13, no. 18, p. 3007, 2023.
- [20] H. Ting and M. Liu, "Multimodal transformer of incomplete MRI data for brain tumor segmentation," *IEEE journal of biomedical and health informatics*, vol. 28, no. 1, pp. 89-99, 2023.
- [21] M. S. Ullah, M. A. Khan, H. M. Albarakati, R. Damaševičius, and S. Alsenan, "Multimodal brain tumor segmentation and classification from MRI scans based on optimized DeepLabV3+ and interpreted networks information fusion empowered with explainable AI," *Computers in biology and medicine*, vol. 182, p. 109183, 2024.
- [22] S. Muksimova, J. Baltaev, and Y. I. Cho, "Brain Tumor Segmentation with Contextual Transformer-Based U-Net," *Electronics*, vol. 15, no. 4, p. 782, 2026.
- [23] J. Xing and J. Zhang, "Segmentation of Brain Tumors Using a Multi-Modal Segment Anything Model (MSAM) with Missing Modality Adaptation," *Bioengineering*, vol. 12, no. 8, p. 871, 2025.
- [24] T. Zhou and S. Zhu, "Uncertainty quantification and attention-aware fusion guided multi-modal MR brain tumor segmentation," *Computers in Biology and Medicine*, vol. 163, p. 107142, 2023.
- [25] T. Zhou et al., "A reliable framework for brain tumor segmentation via multi-modal fusion and uncertainty modeling," *Information Fusion*, p. 104085, 2025.
- [26] M. Fuchs, C. Gonzalez, and A. Mukhopadhyay, "Practical uncertainty quantification for brain tumor segmentation," in *Medical Imaging with Deep Learning*, 2021.
- [27] D. Goswami, R. Subedi, and S. Chakraborty, "MediVLM: A Vision Language Model for Radiology Report Generation from Medical Images," *Findings of the Association for Computational Linguistics: EMNLP*, vol. 2025, pp. 10287-10304, 2025.
- [28] S. Yang, X. Wu, S. Ge, Z. Zheng, S. K. Zhou, and L. Xiao, "Radiology report generation with a learned knowledge base and multi-modal alignment," *Medical Image Analysis*, vol. 86, p. 102798, 2023.
- [29] M. Varol Arisoy, A. Arisoy, and İ. Uysal, "A vision attention driven Language framework for medical report generation," *Scientific Reports*, vol. 15, no. 1, p. 10704, 2025.