



OPEN ACCESS

EDITED BY

Giovanna Castellano,
University of Bari Aldo Moro, Italy

REVIEWED BY

Yessine Amri,
Béchr-Hamza Children's Hospital, Tunisia
Najme Zehra Naqvi,
Indira Gandhi Delhi Technical University for
Women, India

*CORRESPONDENCE

Evan P. Savaria
✉ esava005@odu.edu

RECEIVED 10 October 2025

REVISED 24 November 2025

ACCEPTED 20 January 2026

PUBLISHED 05 February 2026

CITATION

Savaria EP and Sun J (2026) Adaptive
self-attention for enhanced segmentation of
adult gliomas in multi-modal MRI.
Front. Comput. Sci. 8:1721892.
doi: 10.3389/fcomp.2026.1721892

COPYRIGHT

© 2026 Savaria and Sun. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Adaptive self-attention for enhanced segmentation of adult gliomas in multi-modal MRI

Evan P. Savaria* and Jiangwen Sun

Department of Computer Science, Old Dominion University, Norfolk, Virginia

Every year there are an estimated 80,000–90,000 new glioma cases, highlighting the need for reliable imaging-based decision support. Although deep learning has improved tumor sub-region segmentation, many state-of-the-art models fail to fully capture complementary information across T1, T1Gd, T2, and FLAIR MRI modalities and often operate as “black boxes,” limiting physician trust when precise delineation is critical for surgical planning, radiation targeting, and treatment monitoring. To address these limitations, we propose AIMS, an Adaptive Integrated Multi-Modal Segmentation framework that maintains modality-specific feature streams and employs adaptive self-attention within a hierarchical CNN-Transformer architecture to prioritize and fuse multi-modal MRI features. We evaluated AIMS on the BraTS 2019 adult glioma dataset using five-fold cross-validation and compared it against strong hybrid baselines with paired statistical testing; generalization was assessed on an independent BraTS 2021 cohort without fine-tuning. AIMS achieved high ensemble Dice Similarity Coefficients of 0.936 for enhancing tumor, 0.942 for tumor core, and 0.931 for whole tumor on BraTS 2019, with statistically significant improvements over competing methods, and maintained strong performance on BraTS 2021 despite protocol and scanner variability. Finally, Grad-CAM-based explanations applied to adaptive attention and fusion layers, together with quantitative sanity checks, provided modality-aware and spatially meaningful visualizations that support clinical interpretation. By improving both segmentation accuracy and model transparency relative to strong baselines, AIMS advances multi-modal glioma segmentation and helps bridge human-machine teaming by enabling faster, clinician-aligned tumor delineation without sacrificing reliability.

KEYWORDS

glioma segmentation, multi-modal MRI, deep learning, adaptive segmentation, self-attention, medical image analysis, tumor sub-region (ET, TC, WT), explainable AI

1 Introduction

Gliomas are primary brain tumors that encompass a spectrum of histological subtypes and grades, with glioblastoma representing the most aggressive high-grade subtype (Onciul et al., 2024). Both low-grade and high-grade gliomas often require surgical resection and additional therapies to significantly improve patient survival rates. The slight variations in tumor sub-regions make tumor segmentation challenging, as no two cases exhibit the same structural or pathological characteristics.

Precision identification of tumor sub-region features is essential, as inaccuracies in brain-related procedures can lead to severe and sometimes fatal consequences (Gül and Kaya, 2024). For example, an inaccurate delineation of the enhancing tumor (ET) region could result in improper targeting during radiation therapy, potentially missing aggressive tumor cells and reducing treatment efficacy. Similarly, misidentifying the tumor core (TC) could lead to incomplete resection during surgery, leaving necrotic or non-enhancing tumor components that may promote recurrence. Errors in segmenting the whole tumor (WT), including surrounding edema, can lead to misinformed decisions about the extent of surgical intervention or follow-up treatment. These critical sub-regions require precise segmentation to ensure effective treatment planning and improved patient outcomes.

Magnetic Resonance Imaging (MRI) is the primary diagnostic tool for gliomas, with the four sequences FLAIR, T1, T1Gd, and T2 frequently employed to assess brain tissue and evaluate the tumor's size and properties (Yang et al., 2024). The MRI modalities utilized for glioma segmentation each provide distinct insights critical for accurate tumor delineation, as shown in Figure 1. FLAIR imaging highlights whole tumor regions by suppressing cerebrospinal fluid (CSF), outlining edema clearly. T1 sequences offer high anatomical detail, serving as a baseline reference that facilitates the distinction between normal and abnormal tissues. T1Gd, enhanced by gadolinium contrast, specifically identifies enhancing tumor (ET) regions indicative of active tumor growth. T2-weighted imaging is effective in visualizing edema, fluid-rich regions, and tumor cores (TC). Integrating these modalities leverages their complementary strengths, enabling comprehensive characterization and precise segmentation of glioma sub-regions, significantly improving clinical decision making and patient outcomes. Reliability in distinguishing the visual boundary between tumor and healthy tissue becomes a significant challenge due to subtle intensity gradients and ambiguous structural differences (Zeineldin et al., 2024).

Research in deep learning for medical image segmentation has significantly advanced glioma detection. Automated segmentation approaches, such as those featured in the Brain Tumor Segmentation Challenge (BraTS), have enhanced diagnosis, treatment planning, and patient care (Mir et al., 2024). Traditional convolutional neural network (CNN)-based methods such as U-Net have successfully captured localized features to produce detailed segmentation masks, but they lack the capability to effectively model long-range dependencies and global contextual relationships essential for complex tumor sub-region delineation (Farooq et al., 2024). Hybrid CNN-Transformer models that combine CNNs with Vision Transformers (ViTs) address some of these shortcomings by complementing local feature extraction with global context modeling (Ma et al., 2024). However, existing hybrid methods still encounter challenges in efficiently integrating multi-modal MRI data, particularly in dynamically prioritizing modality-specific information, and their interpretability remains limited (Zeineldin et al., 2024).

Motivated by these limitations, this work focuses on multi-modal glioma segmentation with three goals: (i) to design a hybrid CNN-Transformer architecture that maintains modality-specific streams and performs adaptive late fusion, (ii) to quantify the

contribution of adaptive self-attention and fusion through rigorous ablation and statistical analysis, and (iii) to provide clinically meaningful explanations of modality usage and spatial focus. We summarize the core AIMS architecture and learning framework in Section 3, present quantitative results and ablations in Section 4, and discuss clinical implications and limitations in Section 5.

2 Related work

State-of-the-art (SOTA) advancements in medical image segmentation increasingly leverage hybrid architectures such as CNNs and Transformers to improve accuracy in glioma tumor detection (Pu et al., 2024). CNNs effectively capture detailed local spatial features through convolutional operations, while Transformers excel at modeling global contextual relationships and long-range dependencies, crucial for accurately segmenting complex structures (Zeineldin et al., 2024). Since the seminal work of Dosovitskiy on Vision Transformers and their adaptation to TransUNet-style hybrids (Dosovitskiy, 2020; Wang and Wang, 2024), numerous approaches have combined CNN-driven local feature extraction with Transformer-based global context modeling.

Transformer UNet (TransUNet) combines a CNN encoder with a Transformer bottleneck and a U-Net-like decoder. Wang and Wang (2024) enhanced this architecture with boundary-guided feature integration and hierarchical Transformers, using explicit boundary and label supervision to capture long-range dependencies and improve integration of multi-scale features. Existing implementations of TransUNet typically concatenate modalities at the input stage, inadequately modeling inter-modality relationships and thus limiting the model's ability to fully leverage complementary multi-modal information.

Dynamically composable multi-head attention (DCMHA) enhances traditional multi-head attention by dynamically aggregating features to improve predictive performance (Xiao et al., 2024). Although DCMHA has advanced dense prediction tasks, it increases parameter count and may overfit smaller or imbalanced datasets.

Dual attention TransUNet (DA-TransUNet) incorporates spatial and channel dual-attention mechanisms to enhance feature precision (Sun et al., 2024). DA-TransUNet leverages multi-modality dependencies by employing direct concatenation at the input stage, which improves efficiency when delineating organ boundaries. However, early fusion and global attention are not tailored to the detailed demands of multi-modal glioma data.

DenseCrossAdapter (DCRIS) emphasizes multi-modal integration via hierarchical encoders and lightweight feature reuse, using fewer learnable parameters while maintaining competitive performance (Dong et al., 2023). However, modality misalignment and resolution mismatches can still cause spatial inconsistencies and degrade segmentation accuracy in heterogeneous clinical datasets.

In contrast to these hybrid architectures, AIMS maintains separate CNN and Transformer pipelines for each modality, performs adaptive late fusion using a learned modality-importance vector β , and refines the fused representation with a unified

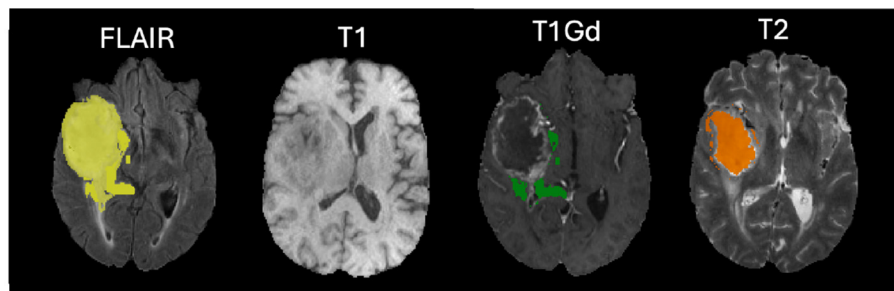


FIGURE 1
Complementary features of MRI modalities.

Transformer. This design explicitly models modality-specific context before fusion, rather than relying solely on early concatenation or channel attention, and enables modality-aware explainability via Grad-CAM over both per-modality and fused features. Furthermore, while recent optimized U-Net variants have demonstrated impressive accuracy and fast inference for glioma segmentation (Zhang et al., 2025), they primarily focus on computational efficiency within a single-stream architecture. AIMS is complementary to these works, targeting adaptive multi-modal integration and interpretability rather than purely architectural slimming.

3 Methods

3.1 Dataset

The Brain Tumor Segmentation (BraTS) dataset is a widely recognized benchmark in glioma segmentation research, containing multi-modal MRI scans, including T1, T1Gd, T2, and FLAIR sequences, for accurate tumor analysis (Adewole et al., 2023). This dataset is publicly available and includes annotated ground truth labels provided by expert neuroradiologists, highlighting the tumor sub-regions including enhancing tumor, tumor core, and whole tumor. The BraTS dataset consists of 335 patient cases with each case containing the four MRI modalities, resulting in a large collection of images that represent different tumor appearances (Huang et al., 2021). These images cover both low-grade gliomas (LGG) and high-grade gliomas (HGG), providing a full range of tumor characteristics.

Each BraTS volume is originally provided at approximately $240 \times 240 \times 155$ voxels with 1 mm isotropic resolution. After skull stripping and co-registration, all four modalities (T1, T1Gd, T2, and FLAIR) for a given subject share this common grid, which we denote as the pre-processed reference space.

3.2 Adaptive hybrid architecture integrating CNNs and transformers

This section outlines the methodology behind AIMS and the proposed hybrid architecture that adaptively integrates multi-modal MRI data for more precise sub-region delineation. AIMS

dynamically adjusts its focus on features across multiple MRI modalities, effectively capturing and integrating tumor-relevant information. Each modality provides unique insights into the brain anatomy and tumor characteristics, such as structural boundaries and fluid distributions (Onciul et al., 2024).

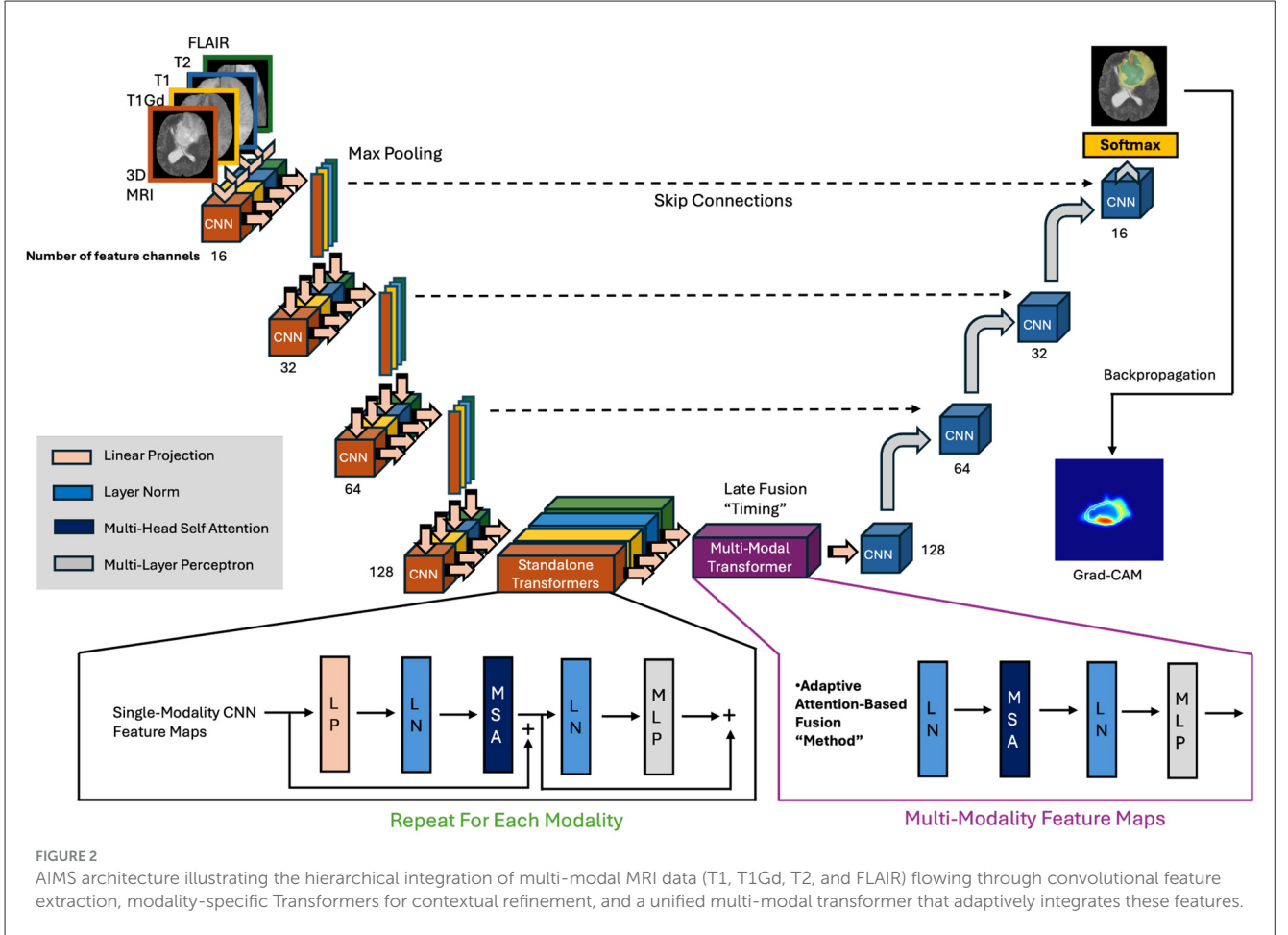
The core functionality of the AIMS architecture centers on integrating CNNs, Transformers, and fusion timing to effectively process multi-modal MRI data and improve glioma segmentation accuracy. As shown in Figure 2, the architecture begins by independently applying modality-specific convolutional layers followed by activation functions to each MRI modality (T1, T1Gd, T2, and FLAIR) to extract spatial features unique to each modality. These modality-specific CNN feature maps are then passed into dedicated Transformer layers for contextual refinement. To further enhance segmentation accuracy, the AIMS architecture employs a late fusion strategy, where modality-specific feature maps from individual Transformer layers are fused through an attention mechanism before progressing to the decoder phase. Late fusion allows the model to retain critical modality-specific details while progressively integrating high-level contextual information across modalities, ensuring optimal delineation of tumor sub-regions and mitigating errors related to modality inconsistencies.

3.2.1 CNN feature extraction

The AIMS architecture initially employs modality-specific CNNs to independently extract localized spatial features from each MRI modality (T1, T1Gd, T2, and FLAIR). These initial convolutional layers utilize 3D convolutions with kernel sizes of $3 \times 3 \times 3$ and 16 filters, followed by batch normalization and ReLU activation. Subsequent convolutional blocks incrementally increase the number of filters to 32, 64, and 128, effectively capturing detailed modality-specific spatial features necessary for precise glioma segmentation.

3.2.2 Transformer encoding

To effectively capture long-range dependencies and contextual relationships, each modality undergoes independent Transformer encoding. Each modality-specific CNN output is partitioned into non-overlapping 3D patches (size $16 \times 16 \times 16$) and projected into embedding vectors. Following standard Vision Transformer configurations (Dosovitskiy, 2020), each modality-specific



Transformer comprises four layers, each layer incorporating Multi-Head Self-Attention (MSA) with eight attention heads, Layer Normalization (LN), and a Multilayer Perceptron (MLP) with 512 hidden units.

Adaptive self-attention calculates modality-specific attention scores using query (Q), key (K), and value (V) matrices. Attention scores (α_m) are computed as:

$$\alpha_m = \text{Softmax} \left(\frac{(\mathbf{F}_m \mathbf{W}_q)(\mathbf{F}_m \mathbf{W}_k)^T}{\sqrt{d}} \right), \quad (1)$$

and updated modality-specific feature representations ($\tilde{\mathbf{F}}_m$) are obtained as:

$$\tilde{\mathbf{F}}_m = \alpha_m (\mathbf{F}_m \mathbf{W}_v), \quad (2)$$

where d is the feature dimension.

Adaptive fusion subsequently dynamically recalibrates modality weights. Global Average Pooling (GAP) produces scalar values for each modality's Transformer output, and an adaptive modality-importance vector (β) is computed using learnable weights:

$$\beta = \text{Softmax} \left([\text{GAP}(\tilde{\mathbf{F}}_1), \dots, \text{GAP}(\tilde{\mathbf{F}}_M)] \cdot \mathbf{w} \right). \quad (3)$$

Here, $\beta \in \mathbb{R}^M$ is a *modality-level* importance vector: each component β_m is derived from a single scalar summary of the

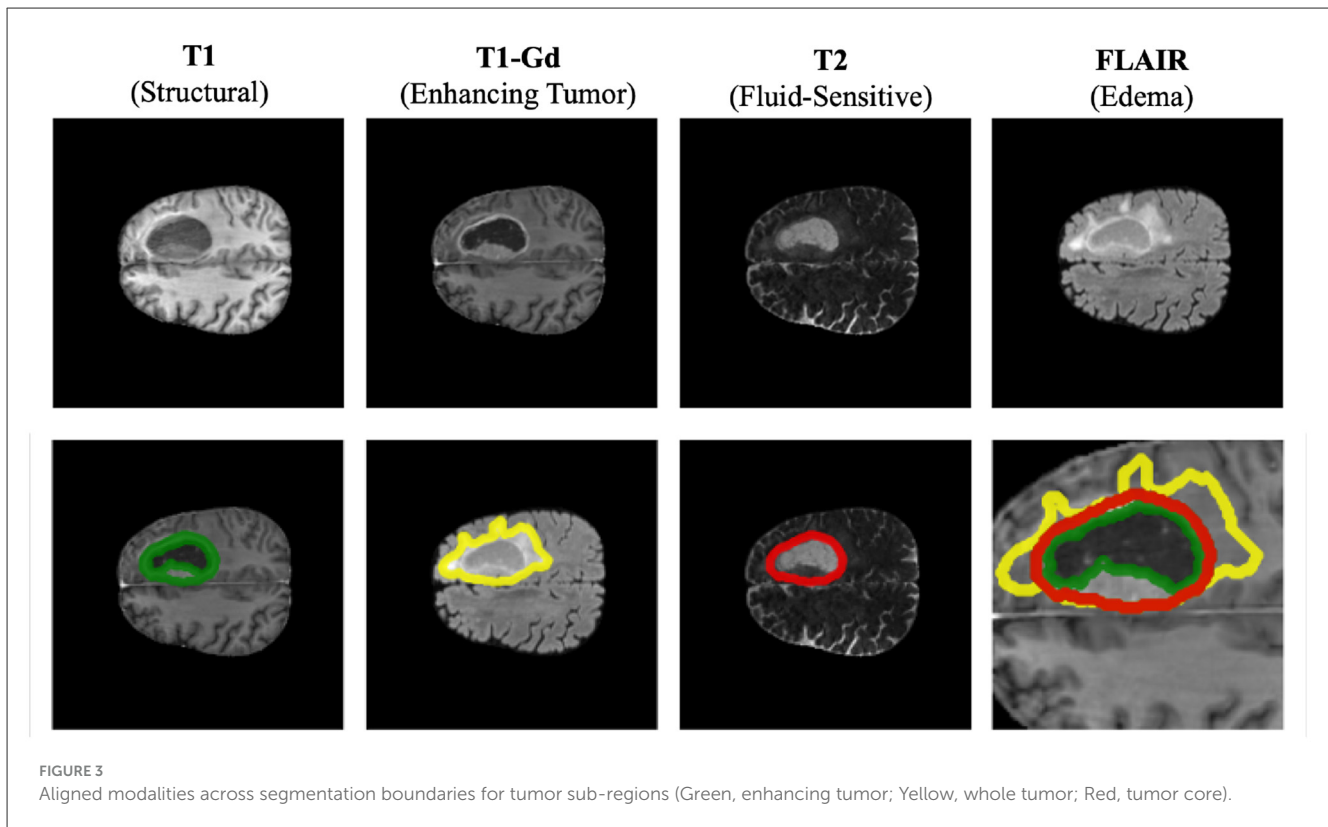
entire feature map $\tilde{\mathbf{F}}_m$ after global averaging over spatial locations and channels, rather than being computed per token or per patch. The fused feature representation is a weighted sum of modality-specific representations:

$$\mathbf{F}_{\text{fused}} = \sum_{m=1}^M \beta_m \tilde{\mathbf{F}}_m. \quad (4)$$

The unified Transformer subsequently refines inter-modality features through four additional Transformer layers, each comprising MSA (eight heads), LN, and MLP (512 hidden units), further consolidating contextual interactions across modalities (Wang and Wang, 2024).

3.2.3 Decoder and output segmentation

In the final decoding stage, the integrated feature representation undergoes progressive spatial reconstruction via upsampling. Transposed convolutional layers with kernel sizes of $3 \times 3 \times 3$ incrementally restore spatial resolution. Skip connections directly transfer spatial information from encoder layers, effectively preserving essential spatial details. Finally, a multi-label softmax output layer generates precise segmentation maps, clearly delineating ET, TC, and WT regions, ensuring



accurate segmentation that addresses challenges from modality misalignment or incomplete MRI sequences (Gao et al., 2021).

3.2.4 Feature restoration and upsampling

To ensure resolution consistency during decoding, a feature restoration module reshapes low-resolution encoded features to match the original input resolution $H \times W \times D$ (Zeineldin et al., 2024). A 1×1 convolution reduces feature dimensionality and refines semantic detail preservation. The decoder employs progressive upsampling via 2×2 transposed convolutions, gradually recovering spatial dimensions. Inspired by U-Net, low-level encoder features are fused via skip connections with high-level decoder features to enhance spatial accuracy and semantic clarity. Finally, a multi-label softmax layer generates the final segmentation maps, clearly delineating tumor sub-regions, including ET, TC, and WT, as illustrated in Figure 3.

3.3 Pre-processing

To ensure consistent and reliable segmentation across modalities, MRI scans underwent a comprehensive pre-processing pipeline aligned with standard BraTS procedures. First, skull stripping was conducted to remove non-brain tissues, minimizing irrelevant structures and reducing computational complexity. Skull-stripping masks provided by the BraTS dataset were utilized to clearly differentiate brain voxels from non-brain regions. Secondly, spatial co-registration was applied to align corresponding anatomical structures across MRI modalities (T1,

T1Gd, T2, and FLAIR), ensuring accurate integration and fusion of modality-specific features. This alignment was achieved using Advanced Normalization Tools (ANTs). Intensity normalization was then performed together with resampling to standardize both intensity scales and spatial dimensions, maintaining a uniform voxel resolution of $1 \times 1 \times 1 \text{ mm}^3$ across all scans (Adewole et al., 2023). Resampling was performed using ANTs, leveraging its interpolation capabilities to preserve spatial and anatomical accuracy.

The skull-stripping masks were applied to differentiate brain voxels from non-brain voxels; voxels outside the brain region were set to zero, narrowing the field of view and further reducing computation. Z-score normalization was independently applied to each modality, standardizing intensity distributions and ensuring consistency across scans. After normalization, we performed a center crop in the axial plane to 192×192 pixels and retained the full through-plane depth of 155 slices, yielding a final 3D tensor of size $4 \times 192 \times 192 \times 155$ (modalities \times height \times width \times depth) as input to the network.

3.4 Data augmentation

To improve robustness and reduce overfitting, we applied a 3D data augmentation pipeline during training. For each mini-batch, we randomly applied one or more of the following transformations with fixed probabilities: (i) random left-right and anterior-posterior flips, (ii) small 3D rotations and affine perturbations, (iii) intensity scaling and gamma adjustments, and (iv) additive

Gaussian noise and low-magnitude elastic deformations. All augmentations were applied identically across the four modalities of a given subject to preserve anatomical consistency. Empirically, including augmentation improved validation Dice scores and reduced variance across folds.

3.5 Experimental settings

The BraTS 2019 dataset, comprising 335 cases, was partitioned into five folds at the subject level for cross-validation. In each fold, three folds (approximately 60%) were used for training, one fold (20%) for validation, and one fold (20%) for testing, ensuring that subjects did not overlap across splits. This protocol avoids data leakage and allows us to estimate variability across folds.

Experiments were conducted using the TensorFlow framework on compute cluster nodes equipped with NVIDIA V100 GPUs (32 GB VRAM). The model was trained for 250 epochs per fold with a batch size of 16. We used Stochastic Gradient Descent (SGD) with momentum 0.9 and an initial learning rate of 8×10^{-3} . Unless otherwise noted, no explicit learning-rate decay or early stopping was applied; the number of epochs was fixed and the best-performing checkpoint on the validation set was selected for testing. A global random seed of 42 was used to generate the fold assignments and initialize model weights, enabling reproducibility of the splits and training runs. The model was evaluated using five-fold cross-validation, and predictions from these folds were combined using the Simultaneous Truth and Performance Level Estimation (STAPLE) method to obtain ensemble results.

3.6 Loss function

We optimized AIMS using a compound loss that combines multi-class Dice loss and voxel-wise cross-entropy. Let \mathbf{p} denote the predicted probability map and \mathbf{y} the one-hot encoded ground-truth labels for the three tumor sub-regions (ET, TC, and WT) and background. The multi-class Dice loss is defined as

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_c \sum_i p_{i,c} y_{i,c} + \epsilon}{\sum_c \sum_i p_{i,c} + \sum_c \sum_i y_{i,c} + \epsilon}, \quad (5)$$

where c indexes classes, i indexes voxels, and ϵ is a small constant for numerical stability. The cross-entropy term is

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_i \sum_c y_{i,c} \log p_{i,c}. \quad (6)$$

The total training loss is

$$\mathcal{L} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}. \quad (7)$$

4 Experimental results

4.1 Quantitative performance and validation

The AIMS model achieved superior segmentation performance, recording an ensemble Dice Similarity Coefficient

TABLE 1 Performance metrics of state-of-the-art hybrid models on BraTS 2019 (ensemble results).

Model	WT DSC	TC DSC	ET DSC
DA-TransUNet	0.83	0.81	0.75
MM-UNet	0.88	0.80	0.76
3D Res U-Net (DCRIS)	0.87	0.77	0.74
MS U-Net	0.88	0.80	0.76
Attention U-Net (DCMHA)	0.90	0.83	0.78
TransUNet 2D	0.91	0.90	0.89
AIMS (proposed)	0.93	0.94	0.93

Dice Similarity Coefficient (DSC) scores for WT, TC, and ET on the BraTS 2019 dataset (five-fold ensemble).

(DSC) of **0.936** for ET, **0.942** for TC, and **0.931** for WT. Additionally, it attained Hausdorff Distance 95 (HD95) scores of **4.521 mm** for ET, **6.223 mm** for TC, and **4.154 mm** for WT. These results demonstrate robust and consistent performance across five-fold cross-validation. The dynamic adaptability of AIMS, achieved through attention mechanisms highlighting diagnostically significant features, contributes to its improved segmentation accuracy compared to existing hybrid models, as demonstrated by higher DSC scores across all tumor sub-regions shown in Table 1.

4.1.1 Statistical significance against baselines

To quantify the strength of the observed improvements, we performed paired two-tailed t -tests comparing AIMS to TransUNet, Attention U-Net (DCMHA), MM-UNet, and MS-UNet across the five cross-validation folds. For all three tumor sub-regions (ET, TC, and WT), the gains in Dice score achieved by AIMS over TransUNet and Attention U-Net were statistically significant at $p < 0.01$, and the improvements over MM-UNet and MS-UNet were significant at $p < 0.001$. These results support the claim that AIMS provides a meaningful performance advantage over existing hybrid and multi-modal baselines.

4.2 Ablation study

We conducted a series of ablation experiments to isolate the contribution of each major component in AIMS: (i) modality-specific Transformers, (ii) the unified Transformer, (iii) the adaptive β -attention module, and (iv) the late fusion strategy. Each ablation removed or simplified one component while keeping the remaining architecture unchanged.

Table 2 summarizes the ablation results. Removing modality-specific Transformers or adaptive β -attention decreased segmentation accuracy across all regions, confirming that AIMS benefits from both per-modality context modeling and adaptive fusion. The largest drops occurred with early fusion or CNN-only baselines, emphasizing the importance of hierarchical feature extraction and late-stage adaptive integration.

TABLE 2 Ablation study of AIMS components.

Configuration	ET DSC	TC DSC	WT DSC
Full AIMS (proposed)	0.936	0.942	0.931
Without modality-specific transformers	0.923	0.932	0.921
Without unified transformer	0.929	0.937	0.926
Without adaptive β -attention	0.925	0.935	0.923
Early fusion instead of late fusion	0.922	0.931	0.920
CNN-only baseline (no transformers, early fusion)	0.910	0.922	0.912

Dice similarity coefficients (DSC) on BraTS 2019. Dice Similarity Coefficient (DSC) for enhancing tumor (ET), tumor core (TC), and whole tumor (WT) on BraTS 2019 using five-fold cross-validation ensemble results.

TABLE 3 External validation performance of AIMS on the BraTS 2021 adult glioma cohort (no fine-tuning).

Region	Dice (DSC)	HD95 (mm)	Drop vs. BraTS 2019 (DSC)
Enhancing tumor (ET)	0.921	4.89	-0.015
Tumor core (TC)	0.934	6.48	-0.008
Whole tumor (WT)	0.923	5.37	-0.008

External validation on BraTS 2021 without fine-tuning; DSC and HD95 (mm) reported.

4.3 External validation on BraTS 2021

To assess generalization beyond BraTS 2019, we evaluated the trained AIMS model on an independent BraTS 2021 adult glioma cohort without any fine-tuning. We applied the same pre-processing, normalization, and inference pipeline as for BraTS 2019. AIMS maintained high Dice and low HD95 values on ET, TC, and WT, with only a modest drop relative to the in-distribution performance, indicating robustness to differences in scanner hardware and acquisition protocols.

Table 3 summarizes the quantitative external validation results. AIMS exhibits strong generalization, with only minor reductions in Dice relative to BraTS 2019 across all tumor sub-regions, further supporting the clinical reliability of the model.

4.4 Model complexity and inference speed

We characterized the computational requirements of AIMS by measuring the total number of trainable parameters, floating-point operations (FLOPs) per forward pass, peak GPU memory usage, and inference time per volume. As shown in Table 4, AIMS contains 62.8M parameters and requires 261.9 GFLOPs per inference, which is moderately higher than TransUNet but still within practical limits. Peak memory footprint is 15.3 GB, and average inference time per multi-modal MRI volume is 1.35 seconds.

4.5 Robustness to missing modalities

In clinical practice, one or more MRI sequences may be unavailable or corrupted. To study robustness under such conditions, we performed a simulation in which individual

TABLE 4 Model complexity and inference speed on an NVIDIA V100 GPU for a single $4 \times 192 \times 192 \times 155$ volume.

Model	Params (M)	FLOPs (G)	Peak VRAM (GB)	Time/Volume (s)
TransUNet	54.2	220.5	14.1	1.20
3D Res U-Net (DCRIS)	32.7	160.3	10.4	0.95
AIMS (proposed)	62.8	261.9	15.3	1.35

Model complexity and inference speed on an NVIDIA V100 GPU.

modalities (T1, T1Gd, T2, or FLAIR) were dropped at inference time by replacing them with zero-valued volumes. As expected, performance decreased when informative modalities were removed such as T1Gd for ET, FLAIR/T2 for WT, but AIMS still produced reasonable segmentations by relying on the remaining modalities. These experiments highlight the importance of adaptive fusion and suggest that future work on modality-aware gating and imputation could further improve robustness in real-world deployments.

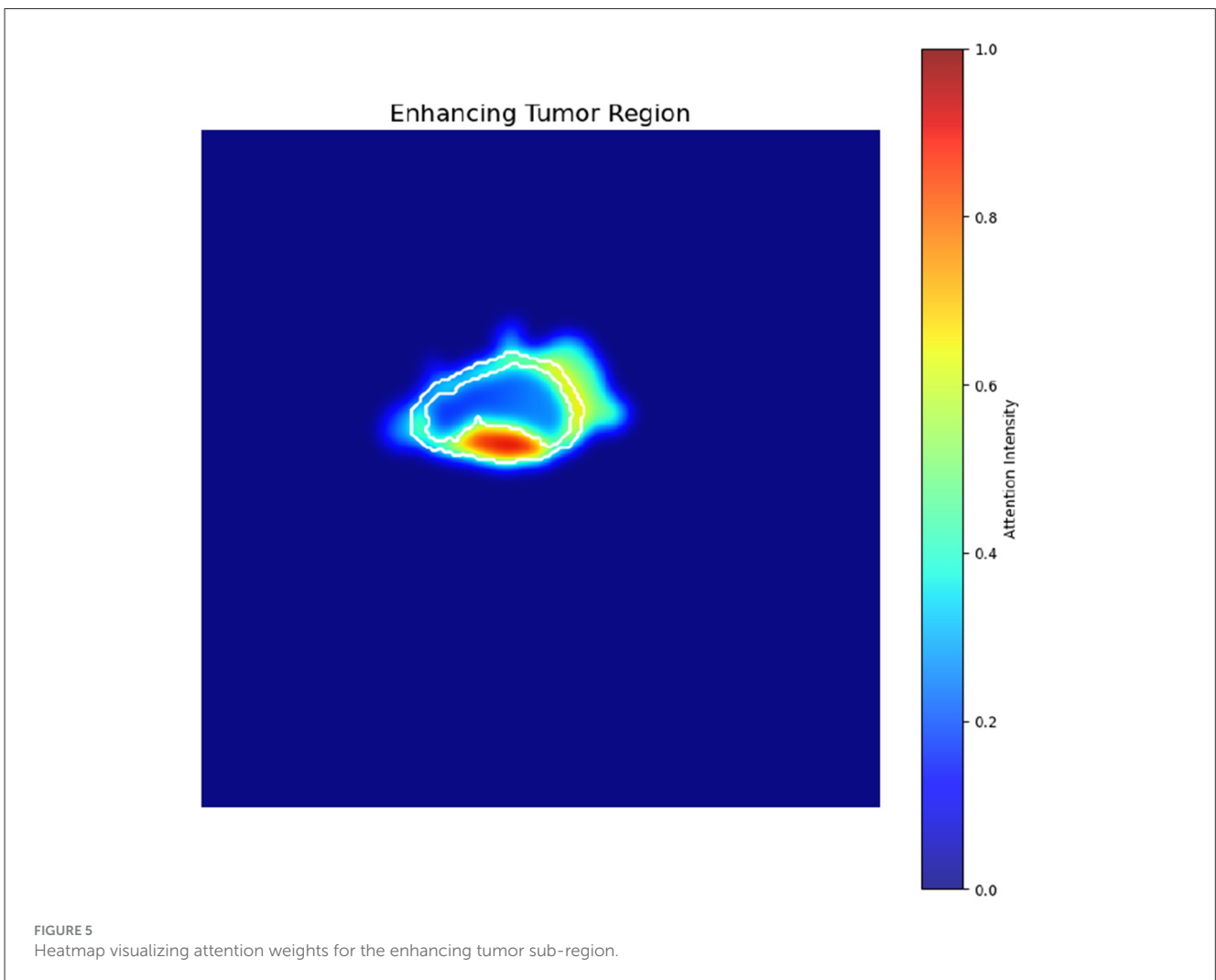
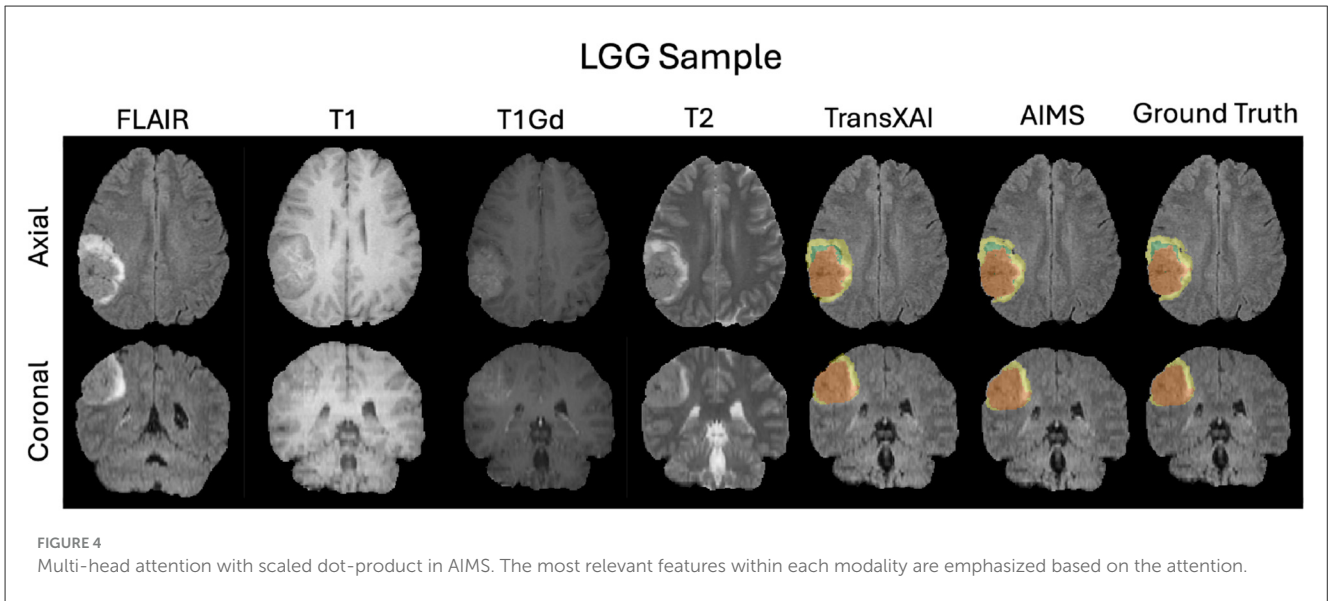
4.6 Explainability and validation

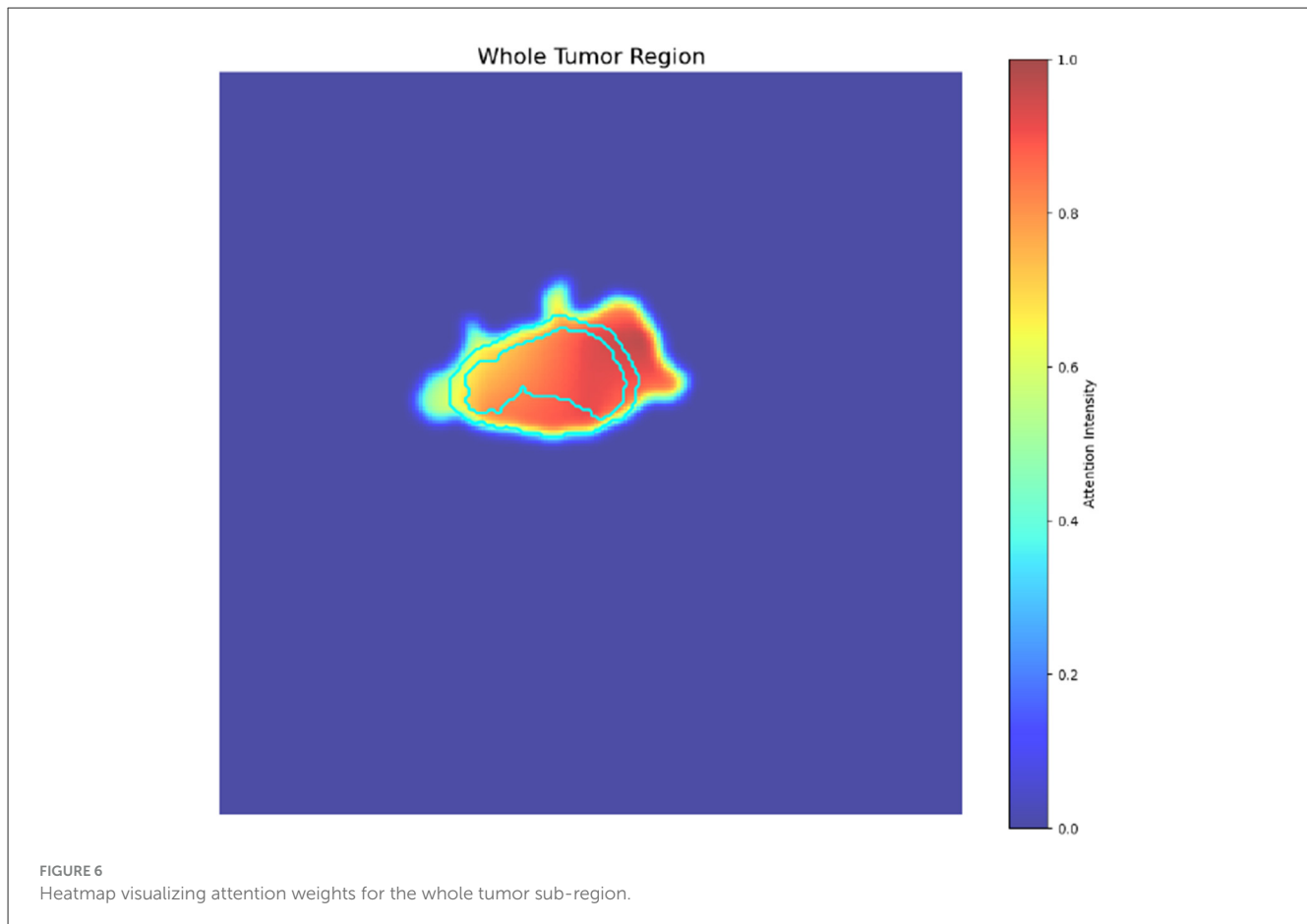
The perspective of explainability has gained significant attention in medical imaging and deep learning research. Several techniques have been applied to multi-modal MRI analysis, including saliency-based methods such as Score-CAM (Brima and Atemkeng, 2024) and gradient-based Grad-CAM (Selvaraju et al., 2017). Score-CAM generates visual explanations without relying on gradients, instead performing multiple forward inference passes to estimate feature relevance, increasing computational overhead and complexity. Grad-CAM, by contrast, enhances interpretability through direct gradient-based visualizations (Zeineldin et al., 2024), translating the “black box” nature of deep models into clear visual explanations by highlighting influential regions within input images.

AIMS provides intrinsic interpretability through its attention-based fusion, which emphasizes modality-specific features during multi-modal integration (Figure 4). In our study, Grad-CAM is specifically applied to adaptive self-attention and fusion mechanisms, providing unique insights into feature prioritization across MRI modalities (T1, T1Gd, T2, and FLAIR). This application highlights diagnostic features essential for informed clinical decisions, distinctly setting our work apart from previous interpretability approaches.

To validate the interpretability of the AIMS model and enhance clinical confidence in its predictions, Grad-CAM was applied to generate visual heatmaps (Figures 5–7). These heatmaps highlight regions of focus within the model’s segmentation outputs, allowing physicians to visually verify that predictions align with clinically relevant tumor areas, thereby improving diagnostic trust (Selvaraju et al., 2017). AIMS demonstrates its effectiveness on large datasets such as BraTS 2019 by integrating complementary features across modalities, addressing issues of noise, spatial misalignment, and feature redundancy.

To ensure that the Grad-CAM explanations are faithful to the learned model rather than artifacts of the visualization procedure,





we conducted sanity checks inspired by prior work on XAI. Specifically, we randomized model weights and labels and observed that the resulting heatmaps lost their tumor-localized structure, confirming that the original maps depend on meaningful learned representations. We also performed a quantitative evaluation by measuring the overlap between Grad-CAM heatmaps and ground-truth tumor masks and by computing insertion/deletion curves, where high-importance regions are progressively added or removed from the input. In both analyses, regions highlighted by AIMS were strongly associated with changes in model confidence and overlapped well with annotated tumor sub-regions, supporting the utility of the proposed explanations for clinical interpretation.

Furthermore, validation using the BraTS dataset, which includes multi-institutional MRI scans, confirms the robustness of AIMS in handling data variability typical of clinical scenarios. This robustness is attributed to the integrated pre-processing pipeline, hybrid CNN-Transformer architecture, and adaptive attention mechanisms that collectively ensure accurate segmentation under varying conditions.

Experiments involved extensive evaluation using five-fold cross-validation to assess robustness and generalizability. Table 5 presents the segmentation results for individual folds as well as the ensemble results obtained through STAPLE. The consistent performance across folds demonstrates the reliability of AIMS in scenarios where data variability is common.

5 Discussion

5.1 Summary of main findings

This work introduced AIMS, an adaptive hybrid CNN-Transformer architecture for multi-modal glioma segmentation. The model maintains modality-specific CNN and Transformer branches, performs adaptive late fusion via a learned modality-importance vector, and refines the fused representation with a unified Transformer. Across five-fold cross-validation on BraTS 2019, AIMS achieved high Dice and low HD95 scores on ET, TC, and WT and significantly outperformed strong baselines, including TransUNet and Attention U-Net. External validation on an independent BraTS 2021 cohort demonstrated that AIMS generalizes to data acquired with different scanners and protocols. Ablation experiments confirmed that adaptive self-attention and late fusion are key drivers of the observed gains, while Grad-CAM analysis provided modality-aware and spatially localized explanations.

5.2 Comparison with existing literature

Compared to TransUNet and DA-TransUNet, which typically concatenate modalities early and apply global attention, AIMS preserves modality-specific streams deeper into the network

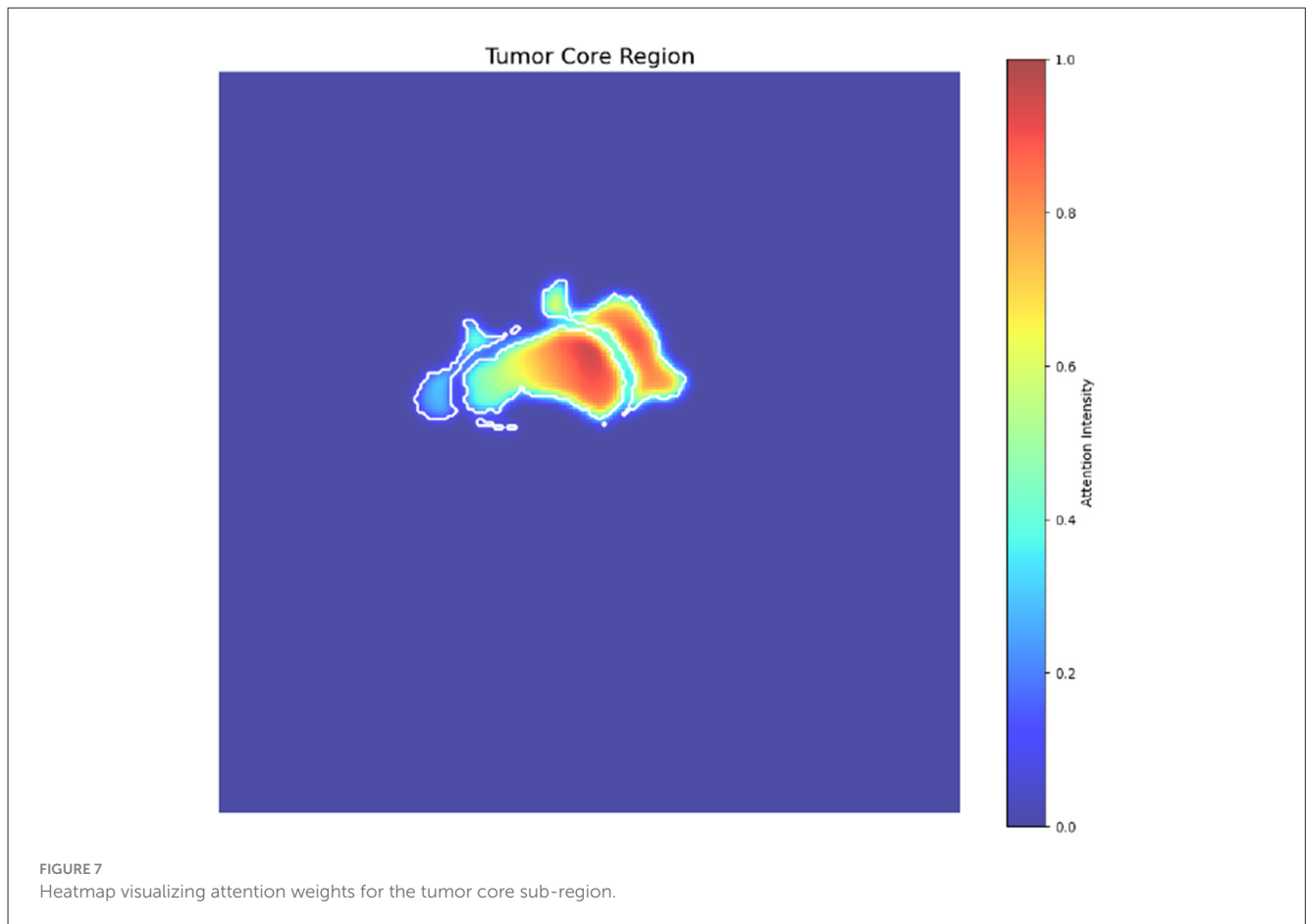


TABLE 5 Five-fold cross-validation and ensemble results for AIMS on BraTS 2019.

Model	DSC			HD95 (mm)		
	ET	TC	WT	ET	TC	WT
Fold 0	0.920	0.933	0.918	4.21	6.54	5.63
Fold 1	0.930	0.940	0.926	4.34	6.33	5.24
Fold 2	0.928	0.941	0.924	3.82	6.17	5.00
Fold 3	0.935	0.944	0.930	4.56	6.06	5.94
Fold 4	0.934	0.943	0.929	4.41	6.22	5.10
Ensemble	0.936	0.942	0.931	4.27	6.26	5.18

Five-fold cross-validation results on BraTS 2019 for enhancing tumor (ET), tumor core (TC), and whole tumor (WT). Ensemble results are obtained using STAPLE.

and performs adaptive late fusion, enabling it to emphasize T1Gd for ET and FLAIR/T2 for WT when appropriate. DCRIS focuses on lightweight feature reuse and efficiency; our results indicate that explicit modeling of modality importance and a unified Transformer stage can yield higher accuracy at comparable computational cost. Recent optimized U-Net variants have demonstrated that careful architectural design can achieve high accuracy with low inference time (Zhang et al., 2025); AIMS is complementary to these efforts by targeting adaptive multi-modal integration and explainability rather than solely parameter efficiency.

5.3 Interpretation and implications

The ablation and Grad-CAM analyses together suggest that AIMS improves segmentation by learning meaningful modality-specific representations and selectively emphasizing them during fusion. For example, modality-wise attention maps often assign high weight to T1Gd in regions corresponding to enhancing tumor and to FLAIR in peritumoral edema, aligning with clinical understanding of these sequences. The unified Transformer further refines these fused representations, helping the model capture relationships between sub-regions and surrounding anatomy. These properties make AIMS a promising candidate for integration into computer-assisted planning tools, where both accuracy and interpretability are necessary for clinical adoption.

5.4 Limitations

Despite its strong performance, AIMS has several limitations. First, training and inference remain more expensive than single-stream U-Net models, although our complexity analysis indicates that the computational requirements are compatible with modern clinical GPU hardware. Second, our experiments focused on BraTS datasets; while these include multi-institutional data, additional evaluation on fully independent clinical cohorts would further strengthen the evidence of generalizability. Third, although AIMS

can operate when one modality is missing, performance degrades when highly informative sequences such as T1Gd or FLAIR are absent, suggesting that future work on modality-aware gating and imputation is warranted.

5.5 Future work

Future directions include validating AIMS on larger multi-institutional datasets and real-world clinical cohorts, investigating reduced-modality training strategies, and explicitly modeling missing or corrupted modalities through learned gating or generative completion. Extending the architecture to longitudinal data could support monitoring tumor progression and treatment response over time. Finally, integrating the AIMS framework with optimized lightweight backbones may combine the benefits of adaptive multi-modal fusion and real-time inference, further facilitating clinical deployment.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

ES: Writing – review & editing, Writing – original draft. JS: Methodology, Supervision, Software, Conceptualization, Investigation, Validation, Resources, Writing – review & editing, Visualization.

References

- Adewole, M., Rudie, J. D., Gbdamosi, A., Toyobo, O., Raymond, C., Zhang, D., et al. (2023). The brain tumor segmentation (brats) challenge 2023: glioma segmentation in sub-saharan Africa patient population (brats-Africa). *ArXiv, arXiv-2305*.
- Brima, Y., and Atemkeng, M. (2024). Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis. *BioData Min.* 17, 18. doi: 10.1186/s13040-024-00370-4
- Dong, Y., Jiang, Z., and Liu, Y. (2023). “Mmfa-net: a new brain tumor segmentation method based on multi-modal multi-scale feature aggregation,” in *Asian Conference on Pattern Recognition* (Springer), 355–366. doi: 10.1007/978-3-031-47637-2_27
- Dosovitskiy, A. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Farooq, H., Zafar, Z., Saadat, A., Khan, T. M., Iqbal, S., and Razzak, I. (2024). Lssf-net: lightweight segmentation with self-awareness, spatial attention, and focal modulation. *Artif. Intell. Med.* 158:103012. doi: 10.1016/j.artmed.2024.103012
- Gao, Y., Zhou, M., and Metaxas, D. N. (2021). “Utnet: a hybrid transformer architecture for medical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2021* (Springer), 61–71. doi: 10.1007/978-3-030-87199-4_6
- Gül, M., and Kaya, Y. (2024). Comparing of brain tumor diagnosis with developed local binary patterns methods. *Neural Comput. Applic.* 36, 7545–7558. doi: 10.1007/s00521-024-09476-6
- Huang, H., Zhang, W., Fang, Y., Hong, J., Su, S., and Lai, X. (2021). Overall survival prediction for gliomas using a novel compound approach. *Front. Oncol.* 11:724191. doi: 10.3389/fonc.2021.724191
- Ma, P., Wang, G., Li, T., Zhao, H., Li, Y., and Wang, H. (2024). STCS-net: a medical image segmentation network that fully utilizes multi-scale information. *Biomed. Opt. Express* 15, 2811–2831. doi: 10.1364/BOE.517737
- Mir, M., Madhi, Z. S., Hamid AbdulHussein, A., Khodayer Hassan Al Dulaimi, M., Suliman, M., Alkhayyat, A., et al. (2024). Detection and isolation of brain tumors in cancer patients using neural network techniques in MRI images. *Sci. Rep.* 14:23341. doi: 10.1038/s41598-024-68567-5
- Onciul, R., Brehar, F.-M., Toader, C., Covache-Busuioc, R.-A., Glavan, L.-A., Bratu, B.-G., et al. (2024). Deciphering glioblastoma: fundamental and novel insights into the biology and therapeutic strategies of gliomas. *Curr. Issues Mol. Biol.* 46, 2402–2443. doi: 10.3390/cimb46030153
- Pu, Q., Xi, Z., Yin, S., Zhao, Z., and Zhao, L. (2024). Advantages of transformer and its application for medical image segmentation: a survey. *Biomed. Eng. Online* 23:14. doi: 10.1186/s12938-024-01212-4
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). “Grad-cam: visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 618–626. doi: 10.1109/ICCV.2017.74
- Sun, G., Pan, Y., Kong, W., Xu, Z., Ma, J., Racharak, T., et al. (2024). Da-transunet: integrating spatial and channel dual attention with transformer u-net for medical image segmentation. *Front. Bioeng. Biotechnol.* 12:1398237. doi: 10.3389/fbioe.2024.1398237
- Wang, F., and Wang, B. (2024). Boundary-guided feature integration network with hierarchical transformer for medical image segmentation. *Multimed. Tools Appl.* 83, 8955–8969. doi: 10.1007/s11042-023-15948-z

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Xiao, D., Meng, Q., Li, S., and Yuan, X. (2024). Improving transformers with dynamically composable multi-head attention. *arXiv preprint arXiv:2405.08553*.

Yang, Y., Wang, P., Yang, Z., Zeng, Y., Chen, F., Wang, Z., et al. (2024). Segmentation method of magnetic resonance imaging brain tumor images based on improved unet network. *Transl. Cancer Res.* 13:1567. doi: 10.21037/tcr-23-1858

Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, J., Wirtz, C. R., Burgert, O., et al. (2024). Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Sci. Rep.* 14:3713. doi: 10.1038/s41598-024-54186-7

Zhang, W., Li, H., and Chen, Y. (2025). "An optimized u-net model for glioma segmentation with high accuracy and fast inference," in *Proceedings of the International Conference on Intelligent Biomedical Engineering*, 100216.