VT INTERNATIONAL PRESS
VTIPJMRR | Journal of Multidisciplinary Research and Review

(Research Article)

Unified Explainable Gated Fusion for Brain Tumor Diagnosis: A Theoretical Framework Extending Cross-Gated Multi-Path Attention Fusion with Multi-Modal Fusion, Segmentation, and Automated Reporting

Vasudha Patil[1], Omkar Joshi[2], Archit Bairagi[3], Prakalp Katla[4], Sarang Chaudhari[5]

[1]*Assistant Professor & HOD, Department of Artificial Intelligence and Data Science, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Maharashtra, India*

[2,3,4,5]*BE Scholar, Department of Artificial Intelligence and Data Science, Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Maharashtra, India*

Corresponding Author: omkarjoshi.13245@gmail.com

Received: 01.09.2026	Accepted: 20.09.2026	Published: 28.09.2026

ABSTRACT

Automated diagnosis of brain tumors from medical images has reached high classification accuracy, but the resulting models rarely localize the lesion, usually accept a single imaging modality, and explain their decisions only post hoc. This paper proposes a theoretical framework that builds on a recently published hybrid of residual convolutional networks and vision transformers joined by cross-gated multi-path attention fusion and trained with a gate-consistency loss. To move from a classifier toward an explainable diagnostic decision-support framework, the model is extended with three components: (a) a masked modality gate that fuses magnetic resonance, computed tomography and positron emission tomography features and remains well defined when some modalities are missing; (b) a shared segmentation decoder; and (c) a traceable explanation and reporting layer. A mathematical formulation is given together with a combined loss in which segmentation is partially supervised, so that samples without masks contribute only to the classification terms. Four properties are proved under stated assumptions: (1) the fusion is a convex combination of the available modality features and is consistent with every availability pattern; (2) branch disagreement is bounded by the gate-consistency penalty; (3) with a linear head, each class score decomposes exactly into location, modality and branch contributions, an algebraic accounting rather than a causal attribution; and (4) integrated-gradient attributions are complete across modalities. An agreement score between the explanation map and the predicted tumor mask, defined for every class including no tumor, flags cases for clinician review. A first-order cost analysis indicates that computation grows linearly with the number of available modalities. No experiments are reported and all claims are theoretical; testable hypotheses and a proof-of-concept protocol on a public dataset with tumor masks are specified for future empirical work.

Keywords: *explainable artificial intelligence; brain tumor classification; multi-modal fusion; vision transformer; gate-consistency loss; missing modalities*

I. INTRODUCTION

Brain and other central nervous system tumors account for a small share of new cancer cases worldwide but a disproportionately large share of cancer deaths [17], so early and accurate diagnosis matters for patient outcomes. Magnetic resonance imaging (MRI) provides detailed information about the lesion, but manual interpretation is time-consuming and partly subjective. Deep learning can support this process, yet clinical use requires more than a class label: the lesion should be localized, the decision should be checkable against other imaging modalities when they are available, and a rationale should accompany the output.

Convolutional neural networks (CNNs), in particular residual networks [2], learn the hierarchical edges and textures that mark tumor borders, but they model long-range context poorly. Vision transformers (ViTs) [3] apply self-attention across image patches and capture global structure, but they can miss subtle local cues and are data-hungry [1], [3]. Hybrid CNN–ViT designs try to balance the two, although most rely on simple concatenation that neither weights nor coordinates the two streams. Muhammad et al. [1] address this with a ResNet–ViT network whose branches are joined by Cross-Gated Multi-Path Attention Fusion (CG-MAF) and trained with a Gate-Consistency Loss (GCL) that ties the predictions of each branch to the fused prediction. That system is, by design, a single-modality classifier: it returns a class label and, after training, post-hoc saliency maps.

 © 2026 VT International Press. All rights reserved. E-ISSN: 3139-7026

This paper argues that the CG-MAF foundation is well placed to grow into an integrated diagnostic decision-support framework, and that the growth can be theoretically coherent rather than a stack of attached modules. The learnable gates that decide how local and global evidence combine are also a natural place to merge several imaging modalities, to feed a segmentation decoder, and to anchor explanations, because every downstream quantity can then be traced to one fused representation. We call the resulting construct the Unified Explainable Gated-Fusion framework (UEGF). The primary study contributes the dual-branch encoder, the gated fusion, and the consistency objective; the proposed layer adds modality gating, tumor segmentation, joint explanation, and automated reporting on top of them, with the modifications of the base model stated explicitly in Section III.

The objectives of this work are: (i) to formalize UEGF as one model with well-defined behavior when imaging modalities are missing; (ii) to specify a composite objective that couples classification, branch consistency, and partially supervised segmentation; (iii) to derive provable properties that link fusion, explanation, and reporting; and (iv) to state testable hypotheses and an evaluation protocol, including a proof of concept on public data, for later empirical work. Because this is a theoretical study, no experimental results are claimed, and statements about diagnostic value are hypotheses rather than findings. The remainder of the paper is organized as follows: Section II reviews related work, Section III presents the framework and its mathematical model, Section IV derives its formal properties, Section V discusses implications and limitations, and Section VI concludes.

II. LITERATURE REVIEW

Research relevant to UEGF can be organized around five questions: how local and global features are combined, how decisions are explained and whether the explanations are faithful, how tumors are delineated, how several modalities are used when some may be missing, and how findings are turned into reports. The review below follows these themes and ends with the gap that UEGF is designed to close.

A. Hybrid Convolution–Transformer Fusion and Gating

Residual networks [2] made very deep CNNs trainable and remain a standard source of local texture features, while the vision transformer [3] treats an image as a sequence of patches and models long-range dependencies through self-attention. Early CNN pipelines for tumor classification, including residual-dense, channel-attention, and lightweight variants surveyed in [1], were strong on fine anatomical detail but limited in global reasoning, and ensemble CNNs gained accuracy at a substantial computational cost. Hybrid models addressed the locality–globality trade-off, yet many fuse features statically, which constrains how the relevance of each stream can adapt to the input. Selective cross-attention inside a ViT backbone [4] and gated dual-branch attention for tumor segmentation [5] moved toward adaptive fusion, but without a mechanism that forces the two streams to agree. CG-MAF with the gate-consistency loss [1] supplies this: learnable gates decide how much each branch trusts the other, and a gate-weighted divergence penalty aligns each branch's prediction with the fused one.

B. Explainability in Tumor Classification

Explainability research in this domain relies mostly on post-hoc attribution. Grad-CAM [6] weights the final convolutional activations by class gradients to give coarse class-specific heatmaps, and Integrated Gradients [7] attributes a prediction to input pixels along a path from a baseline. Frameworks that combine LIME, Grad-CAM, and SHAP [8], and pipelines that pair a DeepLabV3+ segmentation network with Grad-CAM [9], improve the clinical usefulness of the visual output. In these designs, however, the explanation is generated by a separate stage after the model has been trained, so nothing guarantees that it reflects how the fused evidence was actually formed.

Whether such maps are faithful to the model is itself an open question. Adebayo et al. [23] showed with model- and label-randomization tests that some widely used saliency methods produce nearly the same maps for trained and randomized networks, so visual plausibility does not establish faithfulness, and a benchmark on chest radiographs found that saliency maps localize pathology less reliably than radiologists [24]. These results motivate two choices in UEGF: the explanation is tied algebraically to the fused representation that produces the logit (Proposition 3), and the agreement score Φ is treated as a consistency check rather than as evidence that an explanation, or a diagnosis, is correct.

C. Segmentation and Multi-Modal Learning

Segmentation is dominated by encoder–decoder networks derived from U-Net [10], with attention gates added to suppress irrelevant regions [11]. CNN–Transformer hybrids such as BiTr-UNet [12] and multi-modal models that use multi-scale self- and cross-attention [13] show that global context and cross-modal interaction help delineation. Multi-sequence benchmarks such as BraTS [14] have made multi-modal segmentation a standard task, and overlap-based objectives introduced for volumetric segmentation [15] are widely used to train such networks.

D. Learning with Missing Modalities

© 2026 VT International Press. All rights reserved.

A model trained on a complete set of modalities is not automatically usable when one is absent at inference, and a dedicated line of work addresses this. HeMIS [18] encodes each available modality separately and fuses the encodings through their mean and variance across modalities, so that any non-empty subset yields a fixed-size input, and it is trained with random dropping of modalities. RFNet [19] adds region-aware fusion that weights modalities differently for each tumor sub-region, and mmFormer [20] combines modality-specific encoders with a Transformer that models inter-modal dependencies and with auxiliary regularizers that keep each modality's features useful when others are missing. These methods target segmentation of multi-sequence MRI. UEGF adopts their principle that the fusion operator must be well defined for every availability pattern, together with modality dropout during training, but differs in three respects: the fusion is a learned convex combination produced by a masked softmax (Proposition 1) rather than moment pooling or a fusion Transformer; the same fusion weights serve classification, segmentation, and explanation; and the weights appear in the explanation and the report, so the contribution of each modality is visible to the reader. A direct comparison with [18]–[20] is part of the planned protocol (Section III-F).

E. Automated Report Generation

Automated radiology reporting is usually framed as image-to-text generation. Jing et al. [21] combined co-attention over visual features and predicted tags with a hierarchical recurrent decoder, and R2Gen [22] introduced a memory-driven Transformer that records recurring patterns of report structure. Such generative models produce fluent free text but can state findings that the image does not support, and their sentences cannot be traced to specific model quantities. UEGF takes the opposite trade-off: its report is a deterministic template filled from tensors of the forward pass, which is less expressive than generated prose but makes every statement auditable. In the literature surveyed for this work, classification, segmentation, and explanation are generally produced by separate tools, and the final report is still compiled manually by the radiologist.

F. Research Gap

Taken together, four gaps are visible. First, gated hybrid classifiers such as the base study [1] do not delineate the tumor. Second, methods that handle missing modalities [18]–[20] are designed for segmentation and are not coupled to a gated classifier or to its explanation. Third, explanation is attached after training [6]–[9], and its faithfulness is not guaranteed [23], [24]. Fourth, no single formulation connects the label, the mask, the explanation, and the report so that each can be derived from, and audited against, the same representation; generative report models [21], [22] in particular do not offer this traceability. UEGF is proposed to address these four gaps jointly rather than individually.

III. METHODOLOGY

This study is theoretical. Its method is to take the published CG-MAF and GCL formulation [1] as the starting point, state explicitly where it is modified, introduce each extension as an operator acting on the same fused representation, and then derive the properties that follow. No new dataset is collected and no model is trained here.

A. Research Design

The design is analytical and constructive. Four steps are followed: (1) adopt the base dual-branch encoder, gated fusion, and consistency loss, modifying them only where the extensions require it (token-level fusion, a linear main head, and stop-gradient on the gate weights in the loss, each stated where it is introduced in Sections III-C and III-D); (2) translate each identified gap into a requirement, namely multi-modal input with missing-modality tolerance, tumor delineation, intrinsic explanation, and automated reporting; (3) realize each requirement as a function of one shared fused representation F, so that no component has a private view of the evidence; and (4) prove the resulting properties and specify how they would be tested. The guiding principle is a single source of truth: the label, the mask, the explanation, and the report are all computed from F, which is what makes them mutually auditable.

B. Data Considerations

The framework is data-agnostic, but its three extensions place conditions on data. The base study used 7,153 MRI slices compiled from three public repositories into four classes (glioma, meningioma, pituitary tumor, and no tumor), resized to 224 × 224 with ImageNet normalization, and split so that no subject appears in both training and testing. Such single-modality data can train the classification and consistency terms of UEGF unchanged.

Multi-modal fusion additionally needs paired scans of the same subject that are spatially registered to a common reference frame, and segmentation needs pixel-level masks for at least a subset of the samples. UEGF therefore treats mask supervision as optional through an indicator s defined below. The modality set M is deliberately generic. It can be instantiated as {MRI, CT, PET} where paired collections exist, or as the multiple MRI sequences (for example T1, contrast-enhanced T1, T2, and FLAIR) provided by multimodal benchmarks such as BraTS [14] when paired CT and PET are unavailable. Geometric and intensity augmentation follow the base study.

© 2026 VT International Press. All rights reserved.

Cross-modality preprocessing and weight sharing. MRI, CT, and PET differ in intensity physics (relative signal intensity, Hounsfield units, and standardized uptake values) and in spatial resolution, so sharing ImageNet-pretrained backbones across them is a strong assumption that has to be supported by preprocessing. UEGF therefore normalizes each modality before the shared encoder: percentile clipping (0.5–99.5%) followed by z-scoring per volume for MRI, a fixed brain window in Hounsfield units rescaled to [0, 1] for CT, and conversion to SUV followed by clipping at a fixed upper value and rescaling for PET. Each modality then passes through its own input stem, a 3 × 3 convolution with modality-specific normalization that maps the slice to the three channels expected by the backbones; optionally, the normalization layers inside the backbones can also be kept modality-specific, while all other backbone weights are shared. Registration is carried out in 3D before slices are extracted: each CT and PET volume is rigidly or affinely registered to the subject's MRI with a mutual-information criterion and resampled to the MRI grid, so that the same slice index refers to the same anatomy, and slice-wise 2D registration is not relied on. These choices are assumptions of the framework, not established results. Two alternatives are kept for the validation study: separate backbones for each modality, which removes the sharing assumption at roughly $|M|$ times the backbone parameters, and partial sharing, in which the first one or two backbone stages are modality-specific.

C. Analysis Procedure: Formal Model

Notation. Let M be the set of modalities and $a_m \in \{0, 1\}$ indicate whether modality m is available, with at least one $a_m = 1$. The output of the input stem for a registered slice of modality m is $x_m \in \mathbb{R}^{3\times224\times224}$. Let Ω be the 14 × 14 grid of ViT-Base/16 patch positions (N = 196 locations), d = 512 the shared embedding width, and $\odot$ the element-wise product. The ResNet-50 [2] and ViT-Base [3] backbones share weights across modalities apart from the modality-specific stems, so the parameter budget stays close to that of the single-modality base model.

Dual-branch encoding. The final ResNet-50 feature map S_m (7 × 7 × 2048 for a 224 × 224 input) is bilinearly upsampled to the 14 × 14 grid Ω, and the ViT patch tokens T_m (196 × 768, class token excluded) are arranged on the same grid. Both are projected to d dimensions with the projections of the base model [1], which UEGF applies at every location instead of to pooled vectors:

$$r'_m(u) = W_r S_m(u) + b_r, \qquad v'_m(u) = W_v T_m(u) + b_v, \qquad u \in \Omega \quad (1)$$

Cross-gated multi-path attention fusion. In the base model [1] the gate that scales the ViT features is computed from the pooled ResNet projection, and the gate that scales the ResNet features from the pooled ViT projection. UEGF keeps this cross structure but replaces each pooled source vector by a cross-attention read-out at every location, so that a gate can depend on where the evidence lies. With $c^r_m(u)$ the ResNet context read at location u and $c^v_m(u)$ the ViT context:

$$c^r_m(u) = \sum_{j\in\Omega} \mathrm{softmax}_j\left(\frac{\langle v'_m(u), r'_m(j)\rangle}{\sqrt{d}}\right) r'_m(j), \qquad c^v_m(u) = \sum_{j\in\Omega} \mathrm{softmax}_j\left(\frac{\langle r'_m(u), v'_m(j)\rangle}{\sqrt{d}}\right) v'_m(j) \quad (2)$$

$$g^{r\to v}_m(u) = \sigma(W_{rv} c^r_m(u) + b_{rv}), \qquad g^{v\to r}_m(u) = \sigma(W_{vr} c^v_m(u) + b_{vr}) \quad (3)$$

$$f_m(u) = g^{r\to v}_m(u) \odot v'_m(u) + g^{v\to r}_m(u) \odot r'_m(u) \quad (4)$$

The gate $g^{r\to v}_m$, computed from ResNet context, scales the ViT features, and $g^{v\to r}_m$, computed from ViT context, scales the ResNet features. When Ω has a single location (N = 1), the softmax in Equation (2) equals one, so $c^r_m = r'_m$, $c^v_m = v'_m$, and Equations (2)–(4) reduce exactly to the pooled fusion rule of [1]. Token-level cross-attention is therefore a modification of CG-MAF that contains the base rule as a special case, not an exact reuse of it; it is introduced because the segmentation decoder and the class activation map require a spatial fused representation. If the gates are held constant, Equation (4) collapses to a fixed weighted sum, so gated fusion also contains static fusion as a special case.

Modality gating. A learnable vector w_μ scores each modality from its spatially pooled feature $\bar{f}_m = (1/N) \sum_u f_m(u)$. A masked softmax converts the scores into modality weights that vanish for absent modalities:

$$s_m = w_\mu^{\mathrm{T}} \bar{f}_m, \qquad \alpha_m = \frac{a_m \exp(s_m)}{\sum_{k\in M} a_k \exp(s_k)} \quad (5)$$

$$F(u) = \sum_{m\in M} \alpha_m f_m(u), \qquad \bar{F} = \frac{1}{N}\sum_{u\in\Omega} F(u) \quad (6)$$

Heads. A linear classification head acts on the pooled fused feature. This is a second departure from [1], whose main head is a two-layer perceptron; a linear head is chosen because it makes the decomposition of Proposition 3 exact. As in [1], one linear auxiliary head is attached to each branch. In UEGF these heads act on the pooled, modality-weighted branch features $\bar{r} = (1/N) \sum_{u\in\Omega} \sum_{m\in M} \alpha_m r'_m(u)$ and $\bar{v} = (1/N) \sum_{u\in\Omega} \sum_{m\in M} \alpha_m v'_m(u)$, giving the auxiliary logits z_r and z_v:

$$z = W_c \bar{F} + b_c, \quad z_r = A_r \bar{r} + a_r, \quad z_v = A_v \bar{v} + a_v, \qquad p = \mathrm{softmax}(z), \quad p_r = \mathrm{softmax}(z_r), \quad p_v = \mathrm{softmax}(z_v) \quad (7)$$

© 2026 VT International Press. All rights reserved.

The segmentation decoder D consumes the fused map F and modality-weighted ResNet skip features $S_l = \sum_m \alpha_m S_{m,l}$ from encoder stage l, in the manner of U-Net [10], optionally with attention gates [11]:

$$\widehat{M} = \sigma\big(D(F, \{S_l\})\big) \quad (8)$$

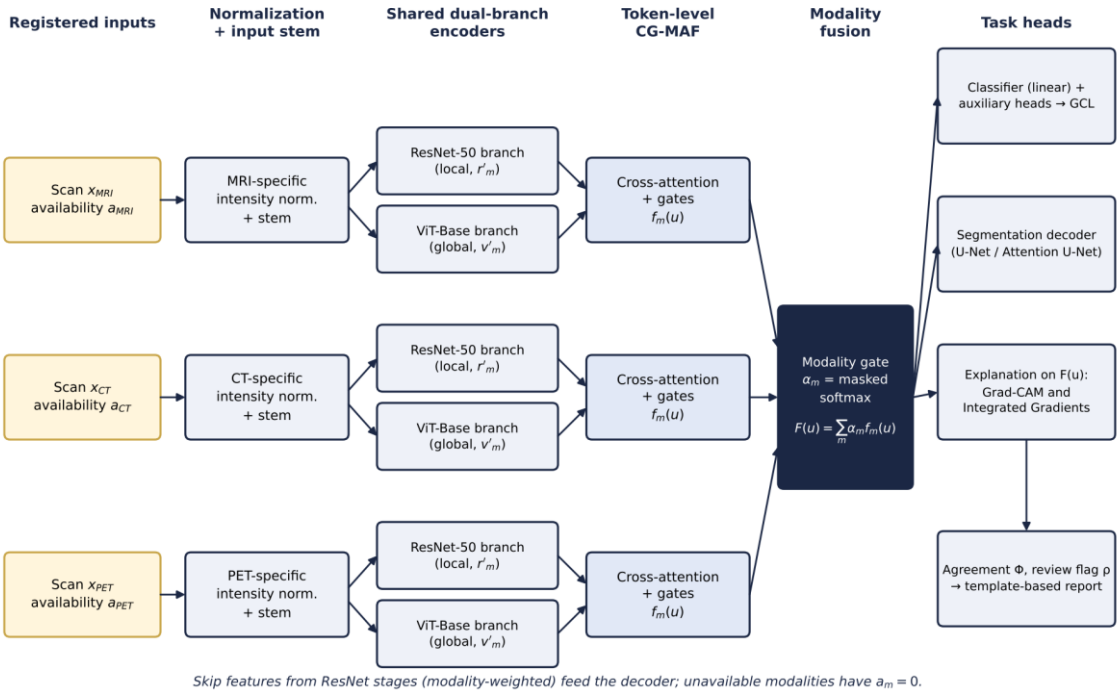


Figure 1. Architecture of the proposed framework. Each available modality is normalized, passed through its own input stem and through the shared ResNet and ViT branches, and fused by token-level CG-MAF; a masked modality gate merges the results into one fused representation that feeds classification, segmentation, explanation, and a template-based report.

D. Training Objective

The base consistency loss is retained and extended with a segmentation term that is active only for samples that carry a mask. With y the one-hot label, KL the Kullback–Leibler divergence, $\bar{g}^{r\to v}$ and $\bar{g}^{v\to r}$ the means of the gates in Equation (3) over modalities, locations, and channels for each sample, and sg(·) the stop-gradient operator:

$$L = -y^{\mathrm{T}} \log p + \frac{1}{2}\mathrm{sg}(\bar{g}^{r\to v})\,\mathrm{KL}(p_v \| p) + \frac{1}{2}\mathrm{sg}(\bar{g}^{v\to r})\,\mathrm{KL}(p_r \| p) + \lambda_s s L_{\mathrm{seg}} \quad (9)$$

Each divergence is weighted by the mean of the gate that scales the corresponding branch in Equation (4): $g^{r\to v}$ multiplies the ViT features v′, so its mean weights the ViT auxiliary term $\mathrm{KL}(p_v \,\|\, p)$, and $g^{v\to r}$ multiplies the ResNet features r′ and weights $\mathrm{KL}(p_r \,\|\, p)$. This is the pairing used in the released implementation of [1], whose consistency weight of 0.5 gives the factor ½. The mean gates enter Equation (9) under stop-gradient, so they act as constants in this term: the consistency penalty cannot be reduced by shrinking the gates, and the gates are trained only through the classification and segmentation terms. In [1] gradients also flow into the gates through the consistency term; stop-gradient is a deliberate change that keeps the bound of Proposition 2 meaningful.

$$L_{\mathrm{seg}} = 1 - \frac{2\sum_u \widehat{M}(u) M(u) + \varepsilon}{\sum_u \widehat{M}(u) + \sum_u M(u) + \varepsilon} + \mathrm{BCE}\big(\widehat{M}, M\big) \quad (10)$$

Here $s \in \{0, 1\}$ marks whether a ground-truth mask M exists, λ_s balances the tasks (default $\lambda_s = 1$, selected from $\{0.25, 0.5, 1, 2\}$ on a validation fold), $\varepsilon = 1$ is a smoothing constant, and the Dice-plus-cross-entropy form follows common practice in volumetric segmentation [15]. Segmentation is therefore partially supervised: samples without a mask contribute nothing to L_{seg}. Because the classification and segmentation gradients both flow into the shared fused map, mask supervision acts as dense auxiliary supervision for the same representation that drives the label.

Modality dropout. To make the model dependable for every availability pattern, each available modality of a training sample is removed independently with probability q (default q = 0.25) by setting its a_m to zero; if no modality remains, the draw is repeated. All modalities are treated alike, so none is privileged as essential, and every non-empty subset of the available modalities occurs with positive probability.

E. Explanation and Reporting

© 2026 VT International Press. All rights reserved.

Explanations are attached to the fused layer, so both branches and all modalities are covered by one map. For class c, with w_c the corresponding row of W_c, Grad-CAM [6] taken at F reduces to the class activation map of Equation (11), including the factor 1/N (Proposition 3), and Integrated Gradients [7] attribute the logit to input pixels relative to a baseline x′:

$$\mathrm{CAM}_c(u) = \frac{\mathrm{ReLU}\left(w_c^{\mathrm{T}} F(u)\right)}{N} \quad (11)$$

$$\mathrm{IG}_i^c(x) = (x_i - x'_i) \int_0^1 \frac{\partial z_c\left(x\prime + t(x - x\prime)\right)}{\partial x_i} dt \quad (12)$$

To test whether the explanation and the delineation tell the same story, an agreement score Φ is computed for the predicted class $\hat{c} = \mathrm{argmax}_c\, p_c$. The CAM is bilinearly upsampled to the mask resolution and divided by its maximum; $B_\tau(\mathrm{CAM})$ keeps the locations where the normalized CAM is at least τ, and $B(\hat{M})$ keeps the locations where $\hat{M}(u) \geq 0.5$. For a tumor class, Φ is the Dice overlap of the two regions and is set to zero if either region is empty. For the no-tumor class, where no CAM region is expected, Φ records whether the decoder also finds no lesion, with $a_{\min}$ a small area tolerance:

$$\Phi = \begin{cases} \frac{2\left|B_\tau(\mathrm{CAM}_{\hat{c}}) \cap B(\hat{M})\right|}{\left|B_\tau(\mathrm{CAM}_{\hat{c}})\right| + \left|B(\hat{M})\right|}, & \text{if } \hat{c} \text{ is a tumor class and both regions are non-empty;} \\ 0, & \text{if } \hat{c} \text{ is a tumor class and either region is empty;} \\ 1\left[\left|B(\hat{M})\right| \leq a_{\min}\right], & \text{if } \hat{c} = \text{ no tumor} \end{cases} \quad (13)$$

$$\rho = 1[\max_c p_c < p_{\min} \text{ or } \Phi < \varphi_{\min}] \quad (14)$$

The flag ρ marks a case for clinician review when the confidence or the agreement is low; a no-tumor prediction is therefore flagged whenever the decoder reports a lesion larger than $a_{\min}$, and a tumor prediction whenever the explanation and the mask disagree. Default values, to be re-selected on a validation fold, are $\tau = 0.5$, $a_{\min} = 0.5\%$ of the image area, $p_{\min} = 0.80$, and $\varphi_{\min} = 0.30$. The selection procedure is fixed in advance: τ is chosen to maximize the mean Φ on correctly classified validation cases, and $p_{\min}$ and $\varphi_{\min}$ are set to the 10th percentiles of $\max_c p_c$ and Φ on those cases, which by the union bound keeps the referral rate among correctly classified validation cases at or below about 20%. The report generator is a deterministic template function of the model outputs: predicted class and probabilities, mask-derived measures such as area and location, the CAM and IG overlays, the modality weights α_m, Φ, and ρ. Because it contains no free-text generation, every field in the report can be traced to a tensor in the forward pass (Figure 2).

F. Planned Validation Protocol

Although empirical results are outside the scope of this paper, the framework is specified so that it can be tested, and the first step is a proof of concept that needs only public data. The Figshare brain tumor dataset of Cheng et al. [25] provides 3,064 contrast-enhanced T1-weighted slices from 233 patients with glioma, meningioma, or pituitary tumor, each with an expert tumor mask, and it overlaps the data of the base study. On this dataset the single-modality form of UEGF ($|M| = 1$) can be trained with subject-level splits and three quantities reported: classification accuracy relative to the base model, the Dice coefficient of the segmentation head, and the distribution of Φ on correctly versus incorrectly classified cases, which tests H2 and H3 directly. Because this dataset has no tumor-free slices, the no-tumor case of Equation (13) requires an additional source, such as the no-tumor images of the base study's composite dataset.

Beyond this proof of concept, classification would be reported with accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices; segmentation with the Dice coefficient and intersection over union, IoU = Dice / (2 − Dice) for a single mask pair; and explanation with Φ and a model-parameter randomization test [23] that checks whether the CAM depends on the trained weights. An ablation ladder would change one component at a time relative to the base CG-MAF + GCL model: pooled versus token-level fusion and perceptron versus linear head; then modality gating (MRI-only versus every non-empty subset of the available modalities, with shared, partially shared, and separate backbones); then the segmentation head; then the agreement analysis. Missing-modality performance would be compared with HeMIS [18], RFNet [19], and mmFormer [20] on BraTS [14]. Splits must remain subject-level, and paired statistical tests across folds should be used, as in the base study.

IV. RESULTS: THEORETICAL ANALYSIS

No experiments are reported in this study, so the results below are formal. They state what the model of Section III guarantees under stated assumptions, compare its capabilities with those of existing system types, and give a first-order account of computational cost. Throughout, the classification head is linear on the pooled fused feature (Equation 7), the modalities are registered to a common grid, and at least one modality is available.

A. Formal Properties

© 2026 VT International Press. All rights reserved.

Proposition 1 (Convex, availability-consistent fusion). Let $\mathcal{A} = \{m : a_m = 1\}$ be non-empty. Then (i) $\alpha_m > 0$ for $m \in \mathcal{A}$, $\alpha_m = 0$ otherwise, and $\sum_m \alpha_m = 1$; (ii) F(u) lies in the convex hull of $\{f_m(u) : m \in \mathcal{A}\}$, hence $\|F(u)\| \le \max_{m\in\mathcal{A}} \|f_m(u)\|$ for any norm; (iii) F and z do not depend on x_m for $m \notin \mathcal{A}$.

Proof. In Equation (5) the numerator is zero when $a_m = 0$ and positive otherwise, and the denominator is a sum of positive terms over $\mathcal{A}$, so (i) holds. Then F(u) in Equation (6) is a convex combination of the available $f_m(u)$, which gives (ii) because norms are convex. For $m \notin \mathcal{A}$ the input x_m enters only through terms multiplied by $a_m = 0$, which gives (iii).

The practical consequence is that one trained head W_c serves all $2^{|M|} - 1$ non-empty availability patterns, and the scale of F does not grow with the number of modalities, as it would under summation. Modality dropout during training (Section III-D) exposes the network to these patterns during learning.

Proposition 2 (Bounded branch disagreement). Let $P_v = \tfrac{1}{2}\,\bar{g}^{r\to v}\,\mathrm{KL}(p_v \,\|\, p)$ and $P_r = \tfrac{1}{2}\,\bar{g}^{v\to r}\,\mathrm{KL}(p_r \,\|\, p)$ be the values of the consistency terms in Equation (9), and TV the total-variation distance. Then $L \ge -y^{T} \log p$, and whenever the mean gates are positive, $\mathrm{TV}(p_v, p) \le \sqrt{(P_v / \bar{g}^{r\to v})}$ and $\mathrm{TV}(p_r, p) \le \sqrt{(P_r / \bar{g}^{v\to r})}$. Consequently $\mathrm{TV}(p_r, p_v) \le \sqrt{(P_v / \bar{g}^{r\to v})} + \sqrt{(P_r / \bar{g}^{v\to r})}$.

Proof. The KL divergence is non-negative, so L is at least the cross-entropy. Pinsker's inequality [16] states $\mathrm{TV} \le \sqrt{(\mathrm{KL}/2)}$, that is, $\mathrm{KL} \ge 2\,\mathrm{TV}^2$. Therefore $P_v \ge \bar{g}^{r\to v}\,\mathrm{TV}(p_v, p)^2$, which rearranges to the first bound; the second is identical, and the third follows from the triangle inequality.

A lower consistency penalty therefore certifies closer agreement between each branch and the fused prediction, and a branch whose features pass through a larger mean gate is held to a tighter bound for the same penalty. The proposition concerns the values of the terms at a given parameter setting, and the bound becomes vacuous as a mean gate approaches zero. Because $\bar{g}$ enters Equation (9) under stop-gradient, the optimizer cannot lower the penalty by shrinking the gates. Whether the gates in fact stay away from zero is not part of the proposition. A heuristic argument suggests that they do, since a vanishing $g^{r\to v}$ removes the ViT contribution from f_m in Equation (4), which would be expected to raise the cross-entropy; this is an expectation to be checked by monitoring $\bar{g}$ during training, not a proven property.

Proposition 3 (Exact additive attribution at the fusion layer). Under the linear head, the class logit decomposes exactly as $z_c = b_c + (1/N) \sum_u \sum_m \alpha_m [\, \kappa^{V}_{m,c}(u) + \kappa^{R}_{m,c}(u) \,]$, where $\kappa^{V}_{m,c}(u) = w_c^{T}(g^{r\to v}{}_m(u) \odot v'_m(u))$ and $\kappa^{R}_{m,c}(u) = w_c^{T}(g^{v\to r}{}_m(u) \odot r'_m(u))$. Moreover, Grad-CAM taken at layer F equals $\mathrm{CAM}^c(u) = \mathrm{ReLU}(w_c^{T} F(u)) / N$.

Proof. Substitute Equations (4) and (6) into $z_c = b_c + w_c^{T} \bar{F}$ and expand, which is linear algebra. For Grad-CAM, $\partial z_c / \partial F_k(u) = w_{c,k}/N$ for every location, so the pooled gradient weight of channel *k* is $w_{c,k}/N$; weighting $F_k(u)$ by it and summing over *k* gives $w_c^{T} F(u)/N$, followed by the rectifier.

This is the central explanatory result. The class evidence splits exactly, as an algebraic identity, into terms carried by the ViT path and by the ResNet path, by each modality, and by each location, and the split is available without perturbation or re-training. It is, however, an accounting of the forward pass and not a causal attribution. The gates in Equation (3) are computed from both branches through cross-attention, and the weights α_m in Equation (5) depend on all available modalities through the softmax, so κ^{V} is not evidence from the ViT alone, κ^{R} is not evidence from the ResNet alone, and the term of one modality is affected by the others. The decomposition shows which path and which modality carry the fused logit, which can suggest whether local texture or global context dominated a decision; whether removing a branch or a modality would change the decision is an interventional question that requires ablation. The constant 1/N, which also appears in Equation (11), disappears after map normalization, so the CAM keeps its usual interpretation. The exactness depends on the linear head and average pooling; a deeper head, such as the perceptron of [1], would make the decomposition approximate.

Proposition 4 (Completeness across modalities). Let z_c be continuous along the straight path from baseline x' to input x and differentiable almost everywhere on it. Then $\sum_i \mathrm{IG}^{c}_{i}(x) = z_c(x) - z_c(x')$. Defining the modality attribution I^{c}_{m} as the sum of IG^{c}_{i} over the pixels of modality *m*, we have $\sum_m I^{c}_{m} = z_c(x) - z_c(x')$, and $I^{c}_{m} = 0$ for every $m \notin \mathcal{A}$ when the baseline matches x in the absent modalities.

Proof. The first identity is the completeness axiom of Integrated Gradients [7], which follows from the fundamental theorem of calculus applied along the path. Summing pixel attributions by modality gives the second. By Proposition 1(iii), z does not depend on x_m for $m \notin \mathcal{A}$, so the partial derivatives with respect to those pixels vanish and so does their attribution.

Completeness makes the modality-level attributions I^{c}_{m} an accounting identity rather than a heuristic. They can be compared with the learned weights α_m to check that a modality the gate trusts is also one on which the logit actually depends; as with Proposition 3, I^{c}_{m} describes dependence along one path from the baseline and is not an interventional effect.

B. Capability Comparison

© 2026 VT International Press. All rights reserved.

Table 1 contrasts four system types by capability. The entries for the base system follow its published description [1], and those for single-purpose systems summarize the literature reviewed in Section II. The last column lists what UEGF provides by construction, as established by the propositions above.

Table 1. Capability comparison of system types (by design, not by measured performance).

Capability	Single-backbone classifier	Segmentation-only network	Base CG-MAF + GCL [1]	Proposed UEGF
Local and global feature fusion	No	Architecture-dependent [12], [13]	Yes	Yes (token-level)
Cross-branch consistency constraint	No	No	Yes	Yes (retained)
Tumor boundary delineation	No	Yes [10]	No	Yes (shared decoder)
Multi-modal input	Seldom	Some models [13], [18]–[20]	No (single modality)	Yes (gated)
Valid when a modality is missing	Not applicable	Dedicated models [18]–[20]	Not applicable	Yes (Prop. 1)
Explanation	Post-hoc, single model	None	Post-hoc Grad-CAM and IG	Exact algebraic decomposition (Prop. 3) plus post-hoc maps
Explanation–mask agreement score	No	No	No	Yes (Φ)
Structured diagnostic report	Manual	Manual	Manual	Automated, traceable

C. Computational Considerations

The base hybrid has 112.00 million parameters and a measured single-image latency of 13.667 ms at 205.3 images per second [1]. These figures were measured for the pooled fusion and perceptron head of [1]. UEGF changes the fusion to token-level cross-attention and adds a decoder, so the base measurements are only a reference point and the estimates below are lower bounds on cost. Because UEGF shares backbone weights across modalities, the parameter count stays close to the base figure, with small additions for the input stems, the modality gate, and the decoder. Computation grows with the number of available modalities $|\mathcal{A}|$, since each one passes through both branches. On the 14×14 grid ($N = 196$, $d = 512$), the per-location projections of Equation (1) need about 0.28 G multiply–accumulate operations per modality, the cross-attention of Equation (2) about $2 \times 2N^2d \approx 0.08$ G, and the gates of Equation (3) about 0.10 G, for a total of roughly 0.46 G, or about 2% of the backbone cost commonly reported for ResNet-50 (≈ 4.1 G) and ViT-B/16 (≈ 17.6 G). The upsampling of the 7×7 ResNet map is negligible. The decoder is the main unknown and may dominate the added cost; it has to be measured. Table 2 gives first-order estimates derived from the base measurements; they are not measurements of UEGF.

Table 2. First-order cost estimate relative to the base hybrid (estimates and bounds, not measurements).

Quantity	Base hybrid (single modality) [1]	UEGF ($\lvert\mathcal{A}\rvert$ available modalities)
Parameters	112.00 M	≈ 112 M shared backbones + stems, gate, decoder (to be measured)
Branch forward passes per study	1	$\lvert\mathcal{A}\rvert$ (batched or sequential)
Latency (sequential, pre-decoder)	13.667 ms	> $\lvert\mathcal{A}\rvert$ × 13.7 ms (lower bound); token-level fusion and decoder extra
Throughput (pre-decoder)	205.3 images/s	< 205 / $\lvert\mathcal{A}\rvert$ studies/s (upper bound)
Fusion cost	$O(d^2)$, pooled vectors	$O(N^2d + Nd^2)$ per modality, ≈ 0.46 G MAC ($N = 196$)
Segmentation decoder	None	Added once per study; cost to be measured

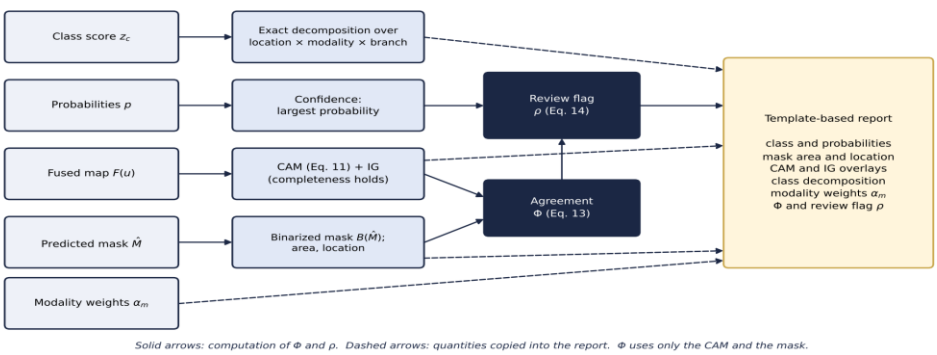


Figure 2. Traceable path from model outputs to the diagnostic report. The agreement score Φ uses only the class activation map and the predicted mask; together with the confidence it sets the review flag. The class decomposition and the modality weights are written to the report but do not enter Φ.

© 2026 VT International Press. All rights reserved.

V. DISCUSSION

The central idea of UEGF is that the gates of CG-MAF are not only a fusion device but also the organizing structure of the whole framework. In the base study they decide how much local detail and global context are trusted. Here the same construct is applied location by location and lifted one level, so that a second gate decides which modality is trusted, and the fused map also feeds the decoder and the explanation. Proposition 1 makes this extension safe for incomplete studies, Proposition 2 keeps the two branches aligned as the task becomes richer, and Propositions 3 and 4 turn the fused representation into an auditable, though not causal, accounting of evidence. The report generator then only reads from quantities that already exist.

The base study reports a mean five-fold accuracy of 99.59% ± 0.40% and a best-run accuracy of 99.81% on its composite single-modality dataset. At that level there is little headroom for accuracy gains, so the value of the extension lies elsewhere: in delineation, in robustness to missing modalities, and in traceability. This framing also sets the correct hypotheses for later experiments, which are stated here as testable propositions rather than results:

H1. Adding CT and PET features through the modality gate improves robustness when MRI quality is degraded, and accuracy degrades gracefully across availability patterns. **H2.** Training with the segmentation term does not reduce, and may improve, classification performance because it supplies dense supervision to the shared fused map. **H3.** The agreement score Φ is higher on correctly classified cases than on misclassified ones, so that ρ identifies cases that merit review. **H4.** The modality weights α_m correlate with the modality attributions I^c_m of Proposition 4.

Several limitations apply. First, this is a theoretical study: none of the hypotheses has been tested, the framework has not been trained, and it should not be read as a validated diagnostic system; the properties hold only under the stated assumptions, notably a linear head with average pooling and registered inputs. Second, UEGF departs from the base model in its token-level fusion, linear head, and stop-gradient on the gate weights, so the accuracy and cost reported in [1] do not transfer automatically. Third, sharing backbones across MRI, CT, and PET relies on modality-specific normalization and stems and on accurate 3D co-registration; if these prove insufficient, separate or partially shared backbones will be needed. Fourth, paired MRI, CT, and PET data with masks are scarce, and the commonly used single-modality classification data carry neither pairing nor masks, so the multi-modal and segmentation components may need multi-sequence benchmarks or institutional data for validation. Fifth, Φ is a consistency indicator, not a calibrated probability of error; a CAM that overlaps the mask does not prove that the diagnosis is correct, and saliency maps can look plausible without being faithful [23]. Sixth, the generated report is a decision-support summary that requires clinician review; clinical use would also require validation across institutions and scanners, which the base study identifies as future work.

VI. CONCLUSION

This paper proposed UEGF, a theoretical framework that extends cross-gated multi-path attention fusion and gate-consistency learning with modality gating, a shared segmentation decoder, and a traceable explanation-and-reporting layer. All components are functions of one fused representation, which allowed four properties to be proved: availability-consistent fusion as a convex combination, bounded branch disagreement, an exact algebraic decomposition of class evidence by location, modality, and branch, and completeness of attribution across modalities. A capability comparison and a first-order cost analysis place the framework relative to existing system types. The contribution is conceptual; its diagnostic value remains to be established empirically.

Future work should first carry out the proof-of-concept study of Section III-F on the public Figshare dataset [25], then test hypotheses H1–H4 on registered multi-modal data with masks, compare missing-modality behavior with [18]–[20], extend the model from 2D slices to 3D volumes, calibrate the agreement score against clinician judgment, and examine deeper heads for which the decomposition is approximate. Deployment-oriented directions include integration with hospital image-archiving and electronic health record systems, cloud and mobile delivery, and domain-shift evaluation across institutions. The authors encourage replication and extension of this study.

ACKNOWLEDGMENT

The authors thank the Principal of Shri Chhatrapati Shivaji Maharaj College of Engineering, Nepti, Dr. Y. R. Kharde, and the project coordinator, G. P. Joshi, for their support, and the Department of Artificial Intelligence and Data Science for providing the facilities used in this work.

REFERENCES

[1] Muhammad, A., Jin, Q., Musaddiq, M. H., Mir, M. S., Alkanan, M., & Gulzar, Y. (2025, November). Explainable AI for brain tumor classification using cross-gated multi-path attention fusion and gate-consistency loss. IEEE Access, 13, 192731–192746. https://doi.org/10.1109/ACCESS.2025.3629532

© 2026 VT International Press. All rights reserved.

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations. https://arxiv.org/abs/2010.11929

[4] Khaniki, M. A. L., Mirzaeibonehkhater, M., Manthouri, M., & Hasani, E. (2025, June). Vision transformer with feature calibration and selective cross-attention for brain tumor classification. Iran Journal of Computer Science, 8(2), 335–347. https://doi.org/10.1007/s42044-024-00220-w

[5] Zhou, Z., Yu, X., Huang, W., Qiu, S., Humayoo, M., Li, H., & Shi, M. (2025, December). Comprehensive exploiting local and global features for brain tumor segmentation: A gated dual-branch hybrid attention mechanism. Biomedical Signal Processing and Control, 110, Article 108285. https://doi.org/10.1016/j.bspc.2025.108285

[6] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017, October). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

[7] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 3319–3328). PMLR.

[8] Akgündoğdu, A., & Çelikbaş, Ş. (2025, October). Explainable deep learning framework for brain tumor detection: Integrating LIME, Grad-CAM, and SHAP for enhanced accuracy. Medical Engineering & Physics, 144, Article 104405. https://doi.org/10.1016/j.medengphy.2025.104405

[9] Saeed, T., Khan, M. A., Hamza, A., Shabaz, M., Khan, W. Z., Alhayan, F., Jamel, L., & Baili, J. (2024, October). Neuro-XAI: Explainable deep learning framework based on DeepLabV3+ and Bayesian optimization for segmentation and classification of brain tumor in MRI scans. Journal of Neuroscience Methods, 410, Article 110247. https://doi.org/10.1016/j.jneumeth.2024.110247

[10] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[11] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. arXiv. https://arxiv.org/abs/1804.03999

[12] Jia, Q., & Shu, H. (2022). BiTr-UNet: A CNN-transformer combined network for MRI brain tumor segmentation. In Proceedings of the BrainLesion Workshop (BrainLes) (pp. 3–14). https://doi.org/10.1007/978-3-031-09002-8_1

[13] Huang, Y., Chen, L., & Zhou, C. (2025). Multi-modal brain tumor segmentation via 3D multi-scale self-attention and cross-attention. arXiv. https://arxiv.org/abs/2504.09088

[14] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., … Van Leemput, K. (2015, October). The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging, 34(10), 1993–2024. https://doi.org/10.1109/TMI.2014.2377694

[15] Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the Fourth International Conference on 3D Vision (pp. 565–571). https://doi.org/10.1109/3DV.2016.79

[16] Cover, T. M., & Thomas, J. A. (2006). Elements of information theory (2nd ed.). Wiley-Interscience.

[17] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229–263. https://doi.org/10.3322/caac.21834

[18] Havaei, M., Guizard, N., Chapados, N., & Bengio, Y. (2016). HeMIS: Hetero-modal image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 (LNCS 9901, pp. 469–477). Springer. https://doi.org/10.1007/978-3-319-46723-8_54

[19] Ding, Y., Yu, X., & Yang, Y. (2021). RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3975–3984). https://doi.org/10.1109/ICCV48922.2021.00394

[20] Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., & Zheng, Y. (2022). mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2022 (LNCS 13435, pp. 107–117). Springer. https://doi.org/10.1007/978-3-031-16443-9_11

[21] Jing, B., Xie, P., & Xing, E. (2018). On the automatic generation of medical imaging reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 2577–2586). https://doi.org/10.18653/v1/P18-1240

[22] Chen, Z., Song, Y., Chang, T.-H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 1439–1449). https://doi.org/10.18653/v1/2020.emnlp-main.112

[23] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In Advances in Neural Information Processing Systems 31 (pp. 9505–9515).

© 2026 VT International Press. All rights reserved.

[24] Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q. H., Nguyen, C. D. T., Ngo, V. D., Seekins, J., Blankenberg, F. G., Ng, A. Y., Lungren, M. P., & Rajpurkar, P. (2022). Benchmarking saliency methods for chest X-ray interpretation. Nature Machine Intelligence, 4(10), 867–878. https://doi.org/10.1038/s42256-022-00536-x

[25] Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., Wang, Z., & Feng, Q. (2015). Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLoS ONE, 10(10), e0140157. https://doi.org/10.1371/journal.pone.0140157

© 2026 VT International Press. All rights reserved.