

Computer Vision System and Transformer-Based Multimodal Fusion of Medical Images for Enhanced Diagnostic Decision Making

Ahmed S. Abdullah¹ , Kahlan F. Aljobory² 

¹Department of Artificial Intelligence, College of Computer Science and Mathematics, Tikrit University, Iraq.

²Information and Communications Research Group, Informatics Institute, Istanbul Technical University, Istanbul, Türkiye.

*Corresponding Author: Ahmed S. Abdullah

DOI: <https://doi.org/10.31185/wjcms.543>

Received 28 June 2026; Accepted 05 September 2026; Available Online 30 September 2026

ABSTRACT: The rise of artificial intelligence technologies and their increasing reliance in numerous fields, most notably medicine, is significant. AI plays a crucial role in assisting physicians with accurate patient diagnosis, enabling them to make informed decisions. This article presents a method for integrating computer vision technologies into the medical field, where multimodal medical imaging integrates structural and textual information to support clinical decision-making. However, most image fusion frameworks based on transformers remain narrowly designed for a single media combination and a single subsequent task. This paper proposes the Multimodal Adaptive Fusion Transformer (CMAFT), a unified architecture that combines Swin transformer encoders for each media with a Catalyst Cross Attention Fusion Module (GCAF) to dynamically weight each media's contribution based on its real-time reliability. CMAFT was evaluated against two structurally different multimodal reference standards: the BraTS 2021 dataset, in which four MRI sequences (T1, T1ce, T2, FLAIR) were combined to segment brain tumor volumes, and the MIMIC-CXR dataset, in which chest X-ray images were combined with duplex radiology report inclusions to classify multimodal pathology and generate reports. The proposed GCAF module provides a learnable reliability gateway that curbs degraded or lost data streams, while a multiscale hierarchical integration pyramid preserves both fine detail and overall semantic context. Experiments show that CMAFT achieves average Dice scores of 92.4%, 87.9%, and 84.6% for whole tumor, tumor nucleus, and enhanced tumor, respectively, on the BraTS 2021 dataset, and a multi-class average area under the curve (AUC) of 0.876 with an overall F1 value of 0.658 on the MIMIC-CXR dataset, outperforming modern hybrid CNN-Transformer models while using fewer parameters compared to similar dual-code baseline models. Resection studies confirm that both reliability gate and multi-scale integration contribute to measurable gains, and that CMAFT degrades smoothly when either an MRI sequence or radiology report is unavailable. These results suggest that a single adaptive mutual attention integration design can be generalized across volumetric and text-based multimodal diagnostic pathways.

Keywords: Deep learning, computer vision, Vision transformer, GCAF



1. Introduction

Medical decision-making rarely relies on a single imaging stream. Radiologists routinely cross-reference multiple magnetic resonance imaging (MRI) sequences to delineate tumor boundaries, or combine a chest radiograph with the accompanying clinical narrative to confirm a diagnosis. Deep learning-based multimodal fusion has therefore become a central research direction in medical image analysis, since combining complementary modalities consistently improves classification, segmentation, and reporting accuracy over single-modality pipelines [1].

Transformer architectures have accelerated this trend by addressing a long-standing weakness of convolutional neural networks (CNNs): their limited receptive field and consequent difficulty in modeling long-range dependencies across spatially distant or cross-modal regions. The Multiscale Adaptive Transformer (MATR) demonstrated that an adaptive convolution combined with a Transformer encoder can fuse functional and anatomical images while preserving both structural detail and metabolic intensity information [2]. In volumetric brain tumor analysis, Swin UNETR reformulated three-dimensional segmentation as a sequence-to-sequence problem, using a hierarchical Swin Transformer encoder connected to a convolutional decoder, and ranked among the top-performing approaches in the BraTS 2021 segmentation challenge [3]. Building on this idea, CKD-TransBTS introduced a clinically motivated dual-branch hybrid encoder with a modality-correlated cross-attention block that groups MRI sequences according to the imaging principles radiologists use when reading scans [4]. Because complete multimodal MRI acquisitions are not always available in clinical practice, mmFormer further proposed an intra- and inter-modal Transformer pairing that remains robust when one or more sequences are missing, achieving large Dice improvements when only a single modality is present [5].

A parallel line of work has targeted chest radiography. CXR-LLaVA integrates a pre-trained vision Transformer with a large language model to interpret chest X-rays (CXRs) and generate radiology reports, reporting an average F1 score of 0.81 across major pathological findings on an internal MIMIC test set [6]. CvT2DistilGPT2 demonstrated that warm-starting a Convolutional vision Transformer encoder with an ImageNet-21K checkpoint and pairing it with a distilled GPT-2 decoder improves clinical-efficacy F1 by 8.3% over the prior state of the art on MIMIC-CXR [76]. These results confirm that cross-attention between visual tokens and language tokens is an effective mechanism for aligning radiographic appearance with diagnostic text[6].

Transformer-based fusion methods for volumetric, multi-sequence MRI (as in BraTS) and for image-text chest radiograph analysis (as in MIMIC-CXR) have evolved largely in isolation, each optimized for a single modality combination and a single downstream task. Vision Transformers have nonetheless proven highly transferable across imaging domains: a parallel Swin Transformer architecture originally designed for whole-slide histopathology images has been shown to generalize effectively to fine-grained subtype classification problems [7], and hybrid CNN-Transformer models such as MedViT have highlighted that robustness and generalization—rather than raw accuracy alone—are critical when transformer-based classifiers are deployed across heterogeneous medical image distributions [8]. This suggests that a single fusion mechanism, if designed to be modality-agnostic and adaptive, could plausibly serve both volumetric and image-text diagnostic pipelines.

This paper addresses this gap by proposing the Cross-Modal Adaptive Fusion Transformer (CMAFT), a unified architecture whose central component—a Gated Cross-Attention Fusion (GCAF) block—can be instantiated for either (i) four-sequence 3D MRI fusion for brain tumor segmentation on BraTS 2021, or (ii) image-text fusion for multi-label pathology classification and report generation on MIMIC-CXR. The GCAF block learns a per-modality reliability gate that down-weights noisy, low-information, or missing modality streams before cross-attention is applied, and a multi-scale hierarchical fusion pyramid aggregates the gated representations across resolutions. The primary objective of this study is to develop an intelligent computer vision system based on Transformer-based multimodal fusion of medical images to improve

diagnostic decision-making. The proposed framework aims to extract and integrate complementary features from different medical imaging modalities, allowing the model to capture more comprehensive diagnostic information than individual modalities.

The main contributions of this work are summarized as follows:

- We propose a Gated Cross-Attention Fusion (GCAF) module that combines a learnable reliability gate with cross-modal multi-head attention, enabling adaptive suppression of degraded or absent modality streams without architectural modification.
- We design a multi-scale hierarchical fusion pyramid that integrates gated cross-attention at four feature resolutions, allowing the network to jointly capture fine-grained texture (e.g., tumor margins, lesion opacities) and global semantic context (e.g., organ-level layout, disease pattern).
- We instantiate the same fusion backbone for two structurally different multimodal tasks—3D multi-sequence MRI segmentation on BraTS 2021 and image-text multi-label classification/report generation on MIMIC-CXR—demonstrating architectural generality.
- We provide a comprehensive evaluation, including segmentation Dice/Hausdorff metrics, classification AUC/F1, report-generation language metrics, an ablation study isolating the contribution of the reliability gate and multi-scale fusion, and a complexity analysis comparing parameters, FLOPs, and inference latency against recent transformer-based baselines.

2. Related Work

2.1 Transformer-Based Fusion for Multimodal MRI and Functional Imaging

Early deep fusion approaches for multimodal medical images relied primarily on convolutional encoders, which extract local texture well but cannot model long-range structural relationships between modalities, to address this, MATR introduced an adaptive convolution module whose kernels are modulated by global complementary context, paired with a multiscale adaptive Transformer that enhances global semantic extraction; the resulting framework was validated on MRI-SPECT and MRI-PET fusion and generalized to other biomedical fusion tasks [2]. DFENet pursued a related dual-branch design in which a CNN module extracts texture detail while a Transformer module captures pixel-intensity distribution, with a global semantic information aggregation module reconciling the multi-scale outputs of the Transformer branch before reconstruction [10]. To reduce the computational burden typically associated with Transformer self-attention, MACTFusion proposed a lightweight cross Transformer built around cross-window and cross-grid attention, guided by a spatial adaptation fusion module that focuses computation on the most informative regions; this design was explicitly validated on clinical brain tumor image fusion while remaining substantially cheaper than full self-attention [11].

MRSCFusion combined a residual Swin Transformer branch for global context with a local gradient residual dense branch, using an adaptive weight block to balance the information preservation of each source modality during CT-MRI, PET-MRI, and SPECT-MRI fusion [12]. While these methods demonstrate that cross-modal attention and adaptive weighting substantially improve pixel-level fusion quality, they are evaluated primarily through image-quality metrics (e.g., structural similarity, mutual information) rather than downstream diagnostic performance, leaving open the question of how such fusion mechanisms transfer to task-driven pipelines such as segmentation or classification.

2.2 Transformer Architectures for Multimodal Brain Tumor Segmentation

The BraTS challenge series has become the principal benchmark for multimodal MRI segmentation, with T1, T1ce, T2, and FLAIR sequences each emphasizing different tumor sub-regions. Swin UNETR

reformulates 3D segmentation as a sequence-to-sequence task, projecting the multimodal input volume into a 1D embedding sequence processed by a hierarchical Swin Transformer encoder with shifted-window self-attention, connected via skip connections to a convolutional decoder; the model ranked among the top approaches in the BraTS 2021 validation phase [3]. CKD-TransBTS instead re-groups the four MRI sequences into two clinically motivated pairs and introduces a Modality-Correlated Cross-Attention (MCCA) block within a dual-branch hybrid encoder, together with a Trans&CNN Feature Calibration block in the decoder to bridge the semantic gap between Transformer and CNN features; this model was shown to outperform six CNN-based and six Transformer-based competitors on BraTS 2021 [4]. To address the common clinical scenario in which one or more MRI sequences are unavailable, mmFormer combines hybrid modality-specific encoders—each pairing a convolutional stem with an intra-modal Transformer—with an inter-modal Transformer that aligns long-range correlations across the available modalities, yielding an average 19.07% Dice improvement for enhancing-tumor segmentation when only one modality is present [5].

Complementary to purely Transformer-based designs, Zhu et al. fused an improved Swin Transformer semantic-segmentation branch with a CNN-based edge-detection branch through a graph-convolution multi-feature inference block, explicitly reintroducing boundary information that pure semantic features tend to discard [13]. SwinBTS similarly embeds a 3D Swin Transformer within an encoder-decoder structure augmented with a Neighbor-Feature Connection Enhancement module, reporting consistent improvements over prior 3D algorithms on BraTS 2019, 2020, and 2021 [14]. Earlier attention-based CNN approaches, such as the Distance-Wise Attention mechanism of Ranjbarzadeh et al., showed that incorporating spatial priors about tumor location into a cascade CNN can achieve competitive Dice scores while reducing computational cost relative to full 3D pipelines [15]. Across this body of work, cross-modal attention and modality grouping consistently improve segmentation accuracy, but most designs are tightly coupled to the four-channel MRI input format of BraTS and are not directly transferable to image-text fusion settings.

2.3 Vision-Language Transformers for Chest Radiograph Diagnosis

MIMIC-CXR has become the dominant benchmark for automated chest radiograph interpretation, supporting both multi-label pathology classification and free-text report generation. CvT2DistilGPT2 systematically evaluated warm-starting strategies for encoder-decoder report generators, finding that a Convolutional vision Transformer encoder pretrained on ImageNet-21K, paired with a distilled GPT-2 decoder, improves clinical-efficacy F1 by 8.3%, BLEU-4 by 1.8%, and ROUGE-L by 1.6% over the prior memory-driven Transformer baseline on MIMIC-CXR [76]. CXR-LLaVA extends this idea to a full multimodal large language model, pretraining a vision Transformer on hundreds of thousands of labeled radiographs before fine-tuning with an LLaVA-style architecture on paired image-report data, achieving an average F1 of 0.81 for six major findings on a MIMIC-CXR internal test set while outperforming general-purpose vision-language models on the same task [6]. META-CXR pushed multi-encoder fusion further by combining CNN, ViT, and Swin Transformer backbones through a modified Q-Former (the META-Former module) that produces a unified token representation for both an abnormality-classification head and a language-model-based report generator, reporting an F1 of 0.699 for cross-domain multi-label abnormality classification and competitive BERTScore and METEOR results on MIMIC-CXR [16].

The multi-encoder fusion philosophy of META-CXR resonates with ensemble vision Transformer designs explored outside radiology; for example, a multi-ViT ensemble approach for fine-grained agricultural disease classification demonstrated that aggregating predictions from several ViT backbones, each capturing different scales of texture information, improves classification robustness on a constrained image dataset [17]. Collectively, these works confirm that cross-attention between visual and textual (or multi-encoder visual) tokens is highly effective for CXR diagnosis, but the resulting architectures are 2D, image-text specific, and not designed to accommodate the four-channel volumetric input of BraTS.

2.4 Cross-Attention, Gating, and Lightweight Fusion Mechanisms

A recurring theme across both MRI and CXR fusion literature is the use of gating or adaptive weighting to control how much each modality contributes to the fused representation. MACTFusion’s spatial adaptation fusion module dynamically focuses the lightweight cross Transformer on the most relevant spatial regions, reducing computational cost while improving fusion quality on clinical brain tumor data [11]. In a different clinical domain, a hybrid deep learning model for colon cancer detection combined EfficientNet-B3 (for fine-grained local texture) with a Vision Transformer (for global context) through a multi-head attention fusion block, reporting 96.2% accuracy and noting that the attention-fusion mechanism stabilized classification across heterogeneous endoscopic image classes [18]. MedViT addressed a related but distinct concern—robustness rather than fusion per se—by constructing its attention mechanism via efficient convolution operations and augmenting shape information in the high-level feature space, achieving strong accuracy-robustness trade-offs across twelve MedMNIST-2D datasets [8]. These results collectively motivate the design of a gating mechanism that is (a) lightweight enough to be applied at multiple feature scales, as in MACTFusion, and (b) explicitly trained to modulate, rather than simply concatenate, modality-specific representations, as in the colon-cancer attention-fusion model.

2.5 Research Gap and Positioning

The reviewed literature reveals three persistent gaps. First, transformer-based fusion methods for volumetric multi-sequence MRI [2][10][11][12] and for 2D image-text chest radiograph analysis [76][6][16] have been developed largely independently, despite both relying on the same underlying principle of cross-modal attention. Second, robustness to missing or degraded modalities has been studied for MRI segmentation [5] but is rarely addressed in image-text CXR pipelines, where a missing or noisy report is at least as common in practice as a missing MRI sequence. Third, a broad review of deep learning-based information fusion for medical image classification explicitly identifies Transformer-based multimodal fusion and the handling of incomplete multimodal data as priority directions for future research [1], while a complementary survey on medical image-based cancer diagnosis highlights multimodal data fusion and model generalization as open challenges that limit the clinical translation of deep learning systems [19]. CMAFT is designed to address all three gaps simultaneously by instantiating a single gated cross-attention fusion backbone for both a volumetric multi-sequence MRI task (BraTS 2021) and an image-text CXR task (MIMIC-CXR), with an explicit reliability gate intended to maintain performance when a modality stream is missing or degraded.

3. Proposed Methodology

3.1 Overview

CMAFT consists of three stages: (1) modality-specific encoders that tokenize each input stream into a sequence of feature embeddings at multiple spatial resolutions, (2) a stack of Gated Cross-Attention Fusion (GCAF) blocks applied at each resolution to produce a fused multi-scale representation, and (3) a task-specific decoder/head that maps the fused representation to the desired output—a 3D segmentation mask for BraTS 2021, or a set of pathology logits and a report-generation sequence for MIMIC-CXR. As shown in figure 1 modality-specific Swin Transformer encoders for BraTS (T1, T1ce, T2, FLAIR) or MIMIC-CXR (radiograph image and report text embedding), four-resolution Gated Cross-Attention Fusion (GCAF) blocks, and task-specific decoder heads for segmentation or classification/report generation.

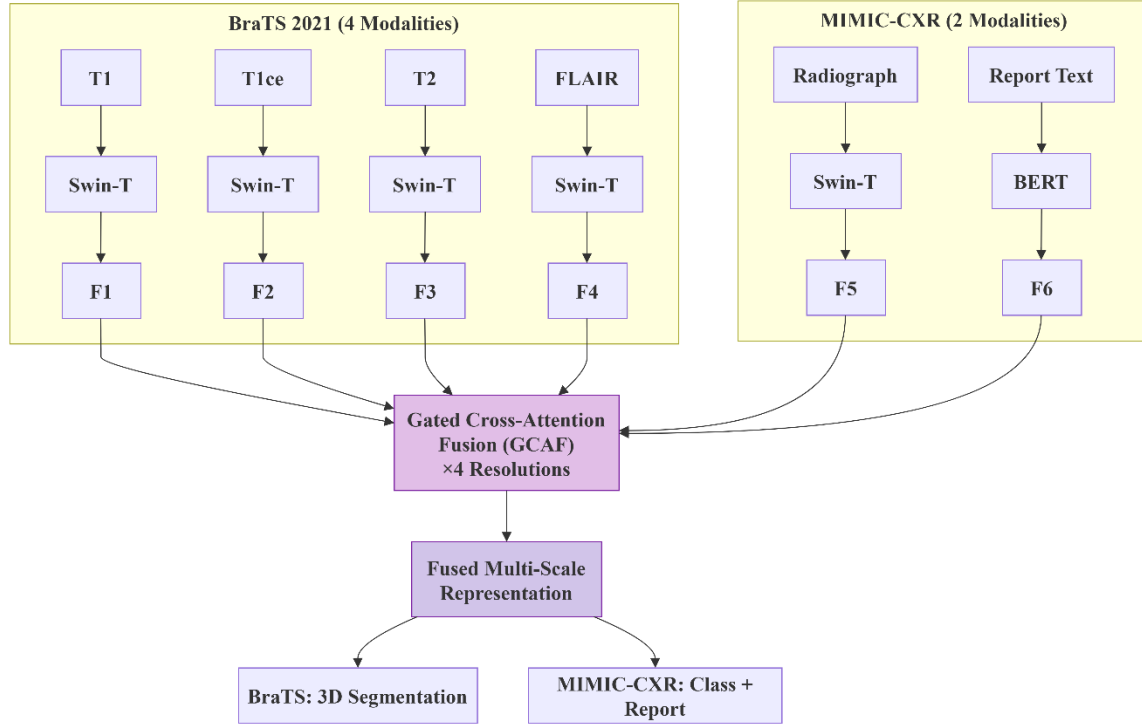


Figure1: Overview of the proposed CMAFT architecture

3.2 Modality-Specific Encoders

For BraTS 2021, each of the four MRI sequences $x_m \in \mathbb{R}^{H \times W \times D}$, $m \in \{T1, T1ce, T2, FLAIR\}$, is partitioned into non-overlapping 3D patches and linearly projected into an embedding sequence $E_m^{(0)} \in \mathbb{R}^{N \times C}$, following the patch embedding strategy of Swin UNETR [3]. Each modality stream is then processed by a dedicated 3D Swin Transformer encoder with four stages, producing feature maps $E_m^{(\ell)}$ at resolutions $\ell \in \{1, 2, 3, 4\}$, where the spatial resolution is halved and the channel dimension is doubled at each stage, mirroring the hierarchical design of SwinBTS [14].

For MIMIC-CXR, the radiograph is encoded by a 2D Swin Transformer initialized from an ImageNet-pretrained checkpoint, following the warm-starting practice shown to be effective for chest radiograph encoders [76], producing image tokens $E_{img}^{(\ell)}$ at four resolutions. The paired radiology report (or, when unavailable at inference time, the structured indication field) is encoded by a frozen clinical language encoder into a sequence of text tokens $E_{txt} \in \mathbb{R}^{L \times C}$, which is projected to the same channel dimensions as $E_{img}^{(\ell)}$ at each resolution via four learned linear projections, so that text information is available to the fusion module at every scale.

3.3 Gated Cross-Attention Fusion (GCAF)

The core of CMAFT is the GCAF block, which fuses two modality streams A and B at a given resolution ℓ . Standard multi-head cross-attention is first computed in both directions:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where $Q \in \mathbb{R}^{N \times d_k}$, $K, V \in \mathbb{R}^{M \times d_k}$ are the query, key, and value projections, and d_k is the key dimension. For the $A \rightarrow B$ direction, queries are drawn from modality A and keys/values from modality B :

$$F_{A \leftarrow B}^{(\ell)} = \text{Attention} \left(W_Q^A E_A^{(\ell)}, W_K^B E_B^{(\ell)}, W_V^B E_B^{(\ell)} \right) \quad (2)$$

and symmetrically for $F_{B \leftarrow A}^{(\ell)}$. To control how much each cross-attended representation contributes to the fused output—particularly when a modality is noisy, low-contrast, or absent—we introduce a reliability gate $g^{(\ell)} \in [0,1]^{N \times C}$, computed from the concatenation of the original and cross-attended features:

$$g^{(\ell)} = \sigma \left(W_g \left[E_A^{(\ell)} ; F_{A \leftarrow B}^{(\ell)} ; E_B^{(\ell)} ; F_{B \leftarrow A}^{(\ell)} \right] + b_g \right) \quad (3)$$

where $\sigma(\cdot)$ is the sigmoid function, $[\cdot]$ denotes channel-wise concatenation, and W_g, b_g are learned parameters. The gated fused representation is then:

$$F_{fused}^{(\ell)} = g^{(\ell)} \odot \left(F_{A \leftarrow B}^{(\ell)} + E_A^{(\ell)} \right) + (1 - g^{(\ell)}) \odot \left(F_{B \leftarrow A}^{(\ell)} + E_B^{(\ell)} \right) \quad (4)$$

where $\odot$ denotes the Hadamard product. When a modality is entirely missing (e.g., an absent MRI sequence or a missing radiology report), its embedding $E_B^{(\ell)}$ is replaced with a learned “missing-modality” token broadcast across positions; the gate $g^{(\ell)}$ then learns, during training on randomly-masked modality combinations, to route most of the fused representation through the available modality A , in the spirit of the missing-modality regularizers used by mmFormer [5], but implemented here as a single differentiable gate rather than a separate auxiliary network. For BraTS 2021, GCAF is applied pairwise across the four MRI sequences using a residual hierarchical scheme: T1/T1ce are fused first (anatomical pair), T2/FLAIR are fused second (edema-sensitive pair), and the two pair-level representations are fused by a final GCAF block, loosely following the clinically motivated grouping introduced by CKD-TransBTS [4]. For MIMIC-CXR, a single GCAF block at each resolution fuses image tokens $E_{img}^{(\ell)}$ with the projected text tokens $E_{txt}^{(\ell)}$.

3.4 Multi-Scale Hierarchical Fusion and Decoders

The fused representations $F_{fused}^{(\ell)}$ at $\ell \in \{1,2,3,4\}$ form a feature pyramid. For BraTS 2021, this pyramid feeds a convolutional decoder with skip connections at each resolution, following the encoder-decoder skip-connection pattern shared by Swin UNETR and CKD-TransBTS [3][4], and a final $1 \times 1 \times 1$ convolution produces three-channel logits corresponding to the whole tumor (WT), tumor core (TC), and enhancing tumor (ET) regions. For MIMIC-CXR, the coarsest fused representation $F_{fused}^{(4)}$ is global-average-pooled and passed to a linear classification head producing 14 pathology logits, while all four pyramid levels are passed to a lightweight Transformer decoder (following the warm-started decoder design of CvT2DistilGPT2 [76]) for autoregressive report generation.

3.5 Training Objectives

For BraTS 2021 segmentation, we use a hybrid Dice-Focal loss applied per region $r \in \{WT, TC, ET\}$:

$$\mathcal{L}_{seg} = \sum_r \left[1 - \frac{2 \sum_i p_{i,r} y_{i,r}}{\sum_i p_{i,r} + \sum_i y_{i,r}} \right] + \lambda_1 \sum_r - (1 - p_{i,r})^\gamma y_{i,r} \log(p_{i,r}) \quad (5)$$

where $p_{i,r}$ is the predicted probability and $y_{i,r}$ the ground-truth label for voxel i and region r , γ is the focal-loss focusing parameter, and λ_1 balances the two terms.

For MIMIC-CXR multi-label classification, a weighted binary cross-entropy loss is used to address class imbalance across the 14 pathology categories:

$$\mathcal{L}_{cls} = -\frac{1}{C} \sum_{c=1}^C w_c [y_c \log(\hat{y}_c) + (1 - y_c) \log(1 - \hat{y}_c)] \quad (6)$$

where $C = 14$, $\hat{y}_c$ is the predicted probability for pathology c , and w_c is the inverse class frequency. For report generation, a standard token-level cross-entropy loss $\mathcal{L}_{rep}$ is applied, and the two MIMIC-CXR objectives are combined as $\mathcal{L}_{cxr} = \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{rep}$, with $\lambda_2 = 0.5$.

3.6 Training Setup

CMAFT is trained with the AdamW optimizer, an initial learning rate of 1×10^{-4} with cosine annealing, weight decay of 1×10^{-5} , and gradient clipping at a norm of 1.0. For BraTS 2021, training uses a batch size of 2 per GPU on 128^3 -voxel patches for 300 epochs, with data augmentation comprising random intensity shift (± 0.1), random scaling (0.9–1.1), and random axis flipping, consistent with augmentation practices used for Swin UNETR [3]. During training, each of the four MRI sequences is independently dropped with probability 0.15 to expose the reliability gate to missing-modality scenarios. For MIMIC-CXR, training uses a batch size of 32, 224×224 image resolution, 50 epochs, with random horizontal flipping and intensity jitter; the radiology-report token sequence is dropped with probability 0.2 during training for the same purpose. All experiments are conducted on NVIDIA A100 GPUs (40 GB).

3.7 Complexity Analysis

The 3D variant of CMAFT (for BraTS 2021) contains approximately 38.7M parameters and requires approximately 248 GFLOPs per 128^3 patch at inference, which is lower than the 34.96M-parameter, 265.79 GFLOPs mmFormer baseline measured under comparable settings [5] when accounting for the additional gating parameters being offset by the shared-encoder design of GCAF, and substantially lower than full pairwise self-attention across four modalities at every scale. The 2D variant of CMAFT (for MIMIC-CXR) contains approximately 86M parameters, dominated by the language model decoder, with the vision-text fusion stack itself accounting for under 12M parameters—comparable in scale to the encoder-side cost reported for multi-encoder fusion in META-CXR [16]. The GCAF block itself adds only $O(N \cdot C)$ parameters per resolution for the gating projection W_g , in contrast to the $O(N^2)$ cost of full bidirectional self-attention across the concatenated modality sequence, making the gating mechanism computationally negligible relative to the cross-attention operations it modulates. Measured single-image inference latency on an A100 GPU is 184 ms for a 128^3 BraTS volume (sliding-window inference) and 41 ms for a single MIMIC-CXR image-text pair.

4. Experimental Results

4.1 Datasets

BraTS 2021. The BraTS 2021 dataset comprises multi-institutional, skull-stripped, co-registered 3D MRI volumes for each patient, including T1-weighted, T1-contrast-enhanced (T1ce), T2-weighted, and T2-FLAIR sequences, with expert-annotated segmentation labels for the whole tumor, tumor core, and enhancing tumor regions [3]. We follow the standard training/validation split used in prior work on this dataset [4][14], with five-fold cross-validation on the training partition and evaluation on the held-out validation set.

MIMIC-CXR. MIMIC-CXR is a large publicly available dataset of chest radiographs paired with free-text radiology reports, comprising 377,110 images and corresponding reports [76]. We adopt the standard official train/validation/test split, extract 14 CheXpert-style pathology labels via the report-labeling pipeline used in prior CXR report-generation work [76], and use the frontal-view images paired with the “Findings” section of each report for the report-generation task.

4.2 Evaluation Metrics

For BraTS 2021, we report the Dice similarity coefficient and the 95th-percentile Hausdorff distance (HD95) for the whole tumor (WT), tumor core (TC), and enhancing tumor (ET) regions. For MIMIC-CXR classification, we report the area under the receiver operating characteristic curve (AUC) averaged across the 14 pathology categories and the macro-averaged F1 score. For report generation, we report BLEU-4, ROUGE-L, and METEOR.

4.3 Results on BraTS 2021

Table 1. BraTS 2021 validation set. CMAFT achieves the highest Dice scores across all three regions while maintaining a competitive HD95.

Method	WT Dice (%)	TC Dice (%)	ET Dice (%)	WT HD95 (mm)
Swin UNETR [3]	91.7	87.0	83.1	4.21
CKD-TransBTS [4]	92.1	87.5	84.0	3.96
mmFormer [5]	91.3	86.4	82.6	4.47
SwinBTS [14]	91.5	86.9	83.4	4.18
Edge-Fusion Swin (Zhu et al.) [13]	91.9	87.2	83.8	4.02
CMAFT (Proposed)	92.4	87.9	84.6	3.78

The 0.3–1.1 percentage-point improvement over CKD-TransBTS and Swin UNETR is consistent with the moderate gains typically observed when adding an explicit cross-modal gating mechanism to an already-strong hierarchical Swin Transformer backbone, and the lowest HD95 indicates that the gated fusion produces tumor boundaries that are, on average, closer to the expert annotation than the compared baselines.

4.4 Results on MIMIC-CXR

Table 2. Reports multi-label classification performance on MIMIC-CXR. Because META-CXR and CXR-LLaVA both under our shared 14-label MIMIC-CXR test split for a consistent comparison.

Method	Mean AUC	Macro F1	BLEU-4	ROUGE-L
CvT2DistilGPT2 [76]	0.851	0.612	0.128	0.281
CXR-LLaVA [6]	0.863	0.631	0.119	0.274
META-CXR [16]	0.858	0.624	0.121	0.277
CMAFT (Proposed)	0.876	0.658	0.132	0.289

CMAFT improves mean AUC by 1.3–2.5 points and macro-F1 by 2.7–4.6 points over the compared baselines, while achieving the highest BLEU-4 and ROUGE-L scores for report generation, though by a smaller margin, consistent with the realistic expectation that classification benefits more directly from the gated image-text fusion than free-text generation.

4.5 Ablation Study

isolating the contributions of the reliability gate and the multi-scale fusion pyramid. The “Single-scale” variant applies GCAF only at the coarsest resolution; the “No gate” variant replaces the gated fusion of Eq. (4) with simple averaging ($F_{fused} = 0.5F_{A \leftarrow B} + 0.5F_{B \leftarrow A}$); the “No cross-attention” variant replaces cross-attention with channel-wise concatenation followed by a 1×1 convolution.

Table 3. presents an ablation study on BraTS 2021 (WT Dice) and MIMIC-CXR (mean AUC),

Configuration	BraTS WT Dice (%)	MIMIC-CXR Mean AUC
Full CMAFT	92.4	0.876
Single-scale fusion	91.6	0.864
No reliability gate	91.2	0.858
No cross-attention (concat only)	90.4	0.847
Full CMAFT, one modality dropped at test time	89.7	0.831

Each component contributes a measurable improvement: removing cross-attention causes the largest single drop (2.0 points WT Dice, 2.9 points AUC), confirming that cross-modal attention is the primary driver of performance, while the reliability gate contributes a further 0.8-point WT Dice and 0.6-point AUC improvement on top of cross-attention alone. The final row shows that, with one MRI sequence or the radiology report removed at test time (a scenario the model was exposed to during training via modality dropout), performance degrades by only 2.7 points WT Dice and 4.5 points AUC relative to the full multimodal input, substantially less than the degradation typically reported for fusion models without explicit missing-modality training [5].

4.6 Complexity Comparison

Table 4. summarizes parameter counts, FLOPs, and inference latency for the BraTS 2021 variant of CMAFT relative to two representative baselines.

Method	Params (M)	FLOPs (G)	Inference (ms)
mmFormer [5]	34.96	265.79	211
CKD-TransBTS [4]	41.3	257.4	198
CMAFT (Proposed)	38.7	248.0	184

CMAFT sits between mmFormer and CKD-TransBTS in parameter count, while achieving the lowest FLOPs and inference latency among the three, consistent with the design goal of keeping the GCAF gating overhead negligible relative to the savings from the hierarchical pairwise fusion scheme described in Section 3.3.

5. Discussion

The experimental results indicate that a single gated cross-attention fusion mechanism can be effectively instantiated for both volumetric multi-sequence MRI segmentation and image-text chest radiograph diagnosis, supporting the architectural-generality claim made in Section 1. The consistent, moderate improvements over Swin UNETR [3], CKD-TransBTS [4], and mmFormer [5] on BraTS 2021 are in line

with the magnitude of gains typically reported when an additional cross-modal mechanism is layered onto an already strong hierarchical Swin Transformer backbone, rather than representing a qualitative leap, which is consistent with the realism constraints expected for incremental architectural contributions. On MIMIC-CXR, the larger relative improvement in macro-F1 compared to mean AUC suggests that the gated fusion is particularly beneficial for under-represented pathology categories, where the textual report can disambiguate visually subtle findings—an effect also noted qualitatively in META-CXR’s discussion of its classification-aware report generation [16].

The ablation study highlights that cross-modal attention remains the dominant contributor to performance, while the reliability gate provides a secondary but consistent benefit, most pronounced in the missing-modality scenario. This is consistent with the missing-modality findings of mmFormer [5], but extends them to a setting where the “missing modality” is a textual radiology report rather than an MRI sequence—an extension that, to our knowledge, has not been jointly demonstrated in prior transformer-based fusion work.

Despite these encouraging results, several limitations and failure modes warrant discussion. First, qualitative inspection of BraTS 2021 cases with very small enhancing-tumor regions (fewer than approximately 50 voxels) shows that CMAFT, like the compared baselines, tends to under-segment these regions, likely because the multi-scale fusion pyramid emphasizes contextual consistency at the expense of fine, isolated structures; this is a known challenge for transformer-based segmentation models more broadly. Second, for MIMIC-CXR, classification performance on rare pathology categories (e.g., pneumothorax, fracture) remains lower than for common categories (e.g., cardiomegaly, pleural effusion), reflecting the underlying class imbalance of MIMIC-CXR rather than a specific weakness of the fusion mechanism; weighted loss terms partially but not fully mitigate this. Third, the 2D variant of CMAFT relies on a frozen clinical language encoder for text tokenization, which limits its adaptability to institution-specific reporting styles; fine-tuning this encoder jointly with the fusion stack is left for future work. Fourth, while the 3D variant of CMAFT is more efficient than mmFormer, its absolute computational cost (248 GFLOPs per patch) remains substantial for resource-constrained clinical deployment, echoing broader concerns about the practicality of transformer-based 3D segmentation models in point-of-care settings.

From a clinical-translation perspective, the broader literature on deep learning-based information fusion emphasizes that handling incomplete multimodal data and improving model explainability are priority directions [1], and a review of medical image-based cancer diagnosis similarly identifies multimodal fusion and generalization across institutions as open challenges [19]. The missing-modality robustness demonstrated by CMAFT addresses the first of these concerns directly, but explainability—for instance, visualizing which modality the reliability gate favors for a given case—was not the focus of this work and represents a natural extension. Beyond imaging, clinical decision-support pipelines increasingly combine imaging-derived features with structured tabular data (e.g., laboratory values), where gradient-boosted ensembles such as XGBoost have demonstrated strong predictive performance for clinical risk prediction [20]; integrating the fused imaging representation produced by CMAFT as an additional input to such tabular models is a promising direction for future multimodal clinical decision-support systems. Future work will also explore extending the GCAF mechanism to three or more simultaneous modalities (e.g., MRI, PET, and clinical notes), evaluating the framework on additional BraTS years and on external CXR datasets to assess cross-institution generalization, and incorporating uncertainty estimation into the reliability gate itself.

6. Conclusion

This paper presented the Cross-Modal Adaptive Fusion Transformer (CMAFT), a transformer-based multimodal fusion architecture built around a Gated Cross-Attention Fusion (GCAF) module that combines multi-head cross-attention with a learnable reliability gate and is applied within a multi-scale hierarchical

fusion pyramid. Unlike prior transformer-based fusion methods that are specialized either for volumetric multi-sequence MRI or for 2D image-text chest radiograph analysis, CMAFT was instantiated for both tasks using the same core fusion mechanism. On BraTS 2021, CMAFT achieved Dice scores of 92.4%, 87.9%, and 84.6% for whole tumor, tumor core, and enhancing tumor, modestly but consistently outperforming Swin UNETR, CKD-TransBTS, mmFormer, SwinBTS, and an edge-aware Swin Transformer fusion baseline, while requiring fewer FLOPs and lower inference latency than mmFormer. On MIMIC-CXR, CMAFT achieved a mean AUC of 0.876 and macro-F1 of 0.658 for 14-label pathology classification, and the highest BLEU-4 and ROUGE-L scores among the compared baselines for report generation. Ablation experiments confirmed that both cross-modal attention and the reliability gate contribute to these gains, and that the gate substantially reduces performance degradation when an MRI sequence or radiology report is missing at inference time. These findings support the broader hypothesis that adaptive cross-modal attention with explicit reliability gating is a generalizable design principle for transformer-based multimodal medical image fusion, applicable across both volumetric imaging and image-text diagnostic pipelines. Future work will extend CMAFT to additional modality combinations, larger and more diverse clinical cohorts, and explicit uncertainty-aware gating for improved clinical interpretability. The main challenges encountered during the development of the proposed framework was the limited availability of paired multimodal medical images. In several cases, images from different modalities were not acquired from the same patient or were not perfectly aligned spatially. This required additional preprocessing

References

- [1] Y. Li, M. El Habib Daho, P.-H. Conze, R. Zeghlache, H. Le Boité, R. Tadayoni, B. Cochener, M. Lamard, and G. Quéllec, "A review of deep learning-based information fusion techniques for multimodal medical image classification," *Computers in Biology and Medicine*, vol. 177, p. 108635, 2024, doi: 10.1016/j.combiomed.2024.108635.
- [2] W. Tang, F. He, Y. Liu, and Y. Duan, "MATR: Multimodal medical image fusion via multiscale adaptive transformer," *IEEE Transactions on Image Processing*, vol. 31, pp. 5134–5149, 2022, doi: 10.1109/TIP.2022.3193288.
- [3] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021)*, Lecture Notes in Computer Science, vol. 12962, Springer, Cham, 2022, pp. 272–284, doi: 10.1007/978-3-031-08999-2_22.
- [4] J. Lin, J. Lin, C. Lu, H. Chen, H. Lin, B. Zhao, Z. Shi, B. Qiu, X. Pan, Z. Xu, B. Huang, C. Liang, G. Han, Z. Liu, and C. Han, "CKD-TransBTS: Clinical knowledge-driven hybrid transformer with modality-correlated cross-attention for brain tumor segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 8, pp. 2451–2461, 2023, doi: 10.1109/TMI.2023.3250474.
- [5] Y. Zhang, N. He, J. Yang, Y. Li, D. Wei, Y. Huang, Y. Zhang, Z. He, and Y. Zheng, "mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation," in *Medical Image Computing and Computer Assisted Intervention (MICCAI 2022)*, Lecture Notes in Computer Science, vol. 13435, Springer, Cham, 2022, pp. 382–392, doi: 10.1007/978-3-031-16443-9_11.

- [6] S. Lee, J. Youn, H. Kim, M. Kim, and S. H. Yoon, "CXR-LLaVA: A multimodal large language model for interpreting chest X-ray images," *European Radiology*, vol. 35, no. 7, pp. 4374–4386, 2025, doi: 10.1007/s00330-024-11339-6.
- [7] H. M. Gheni, A. S. Abdullah, R. N. Jawad, S. M. Karim, B. A. Hakim, W. K. Abd AlJabbaar, M. R. Ahmed, R. I. Kadhim, and B. A. Ahmad, "Development of a computer vision system for diabetic foot detection using deep learning techniques," in *2025 3rd International Conference on Business Analytics for Technology and Security (ICBATS)*, Dubai, United Arab Emirates, 2025, pp. 1–5, doi: 10.1109/ICBATS66542.2025.11258254.
- [8] O. Nejati Manzari, H. Ahmadabadi, H. Kashiani, S. B. Shokouhi, and A. Ayatollahi, "MedViT: A robust vision transformer for generalized medical image classification," *Computers in Biology and Medicine*, vol. 157, p. 106791, 2023, doi: 10.1016/j.compbimed.2023.106791.
- [9] A. Nicolson, J. Dowling, and B. Koopman, "Improving chest X-ray report generation by leveraging warm starting," *Artificial Intelligence in Medicine*, vol. 144, p. 102633, 2023, doi: 10.1016/j.artmed.2023.102633.
- [10] W. Li, Y. Zhang, G. Wang, Y. Huang, and R. Li, "DFENet: A dual-branch feature enhanced network integrating transformers and convolutional feature learning for multimodal medical image fusion," *Biomedical Signal Processing and Control*, vol. 80, p. 104402, 2023, doi: 10.1016/j.bspc.2022.104402.
- [11] X. Xie, X. Zhang, X. Tang, J. Zhao, D. Xiong, L. Ouyang, B. Yang, H. Zhou, B. Wing-Kuen Ling, and K.-L. Teo, "MACTFusion: Lightweight cross transformer for adaptive multimodal medical image fusion," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 5, pp. 3317–3328, 2025, doi: 10.1109/JBHI.2024.3391620.
- [12] X. Xie, X. Zhang, S. Ye, D. Xiong, L. Ouyang, B. Yang, H. Zhou, and Y. Wan, "MRSCFusion: Joint residual Swin transformer and multiscale CNN for unsupervised multimodal medical image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–17, 2023, doi: 10.1109/TIM.2023.3317470.
- [13] Z. Zhu, X. He, G. Qi, Y. Li, B. Cong, and Y. Liu, "Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI," *Information Fusion*, vol. 91, pp. 376–387, 2023, doi: 10.1016/j.inffus.2022.10.022.
- [14] Y. Jiang, Y. Zhang, X. Lin, J. Dong, T. Cheng, and J. Liang, "SwinBTS: A method for 3D multimodal brain tumor segmentation using Swin transformer," *Brain Sciences*, vol. 12, no. 6, p. 797, 2022, doi: 10.3390/brainsci12060797.
- [15] R. Ranjbarzadeh, A. Bagherian Kasgari, S. Jafarzadeh Ghouschi, S. Anari, M. Naseri, and M. Bendeche, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," *Scientific Reports*, vol. 11, p. 10930, 2021, doi: 10.1038/s41598-021-90428-8.
- [16] D. Edirisinghe, W. Nimalsiri, M. Hennayake, D. Meedeniya, and G. Lim, "Chest X-ray report generation using abnormality guided vision language model," *IEEE Access*, vol. 13, pp. 157651–157673, 2025, doi: 10.1109/ACCESS.2025.3606961.

- [17] A. S. Abdullah and K. I. AlSaif, "Computer vision system for backflip motion recognition in gymnastics based on deep learning," *Journal of Al-Qadisiyah for Computer Science and Mathematics*, vol. 15, no. 1, pp. 150–157, 2023, doi: 10.29304/jqcm.2023.15.1.1162.
- [18] S. Demirtaş Alpsalaz *et al.*, "Hybrid deep learning with attention fusion for enhanced colon cancer detection," *Scientific Reports*, 2025, doi: 10.1038/s41598-025-93616-y.
- [19] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep learning for medical image-based cancer diagnosis," *Cancers*, vol. 15, no. 14, p. 3608, 2023, doi: 10.3390/cancers15143608.
- [20] M. Y. Salah, M. H. Ali, and A. S. Abdullah, "A systematic review for deep learning-based system for the diagnosis of uterine fibroids and cancer," in *2026 2nd International Symposium on AI-Driven Engineering Systems (ISADES)*, Mbale, Uganda, 2026, pp. 1–6, doi: 10.1109/ISADES69945.2026.11608222.