International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

DIFFUFUSEMED: DIFFUSION-GUIDED MULTIMODAL TRANSFORMER FOR ROBUST AND CALIBRATED MEDICAL IMAGE SEGMENTATION

[1] S Nandini, [2]M Nagabushanam,[3]G S Nandeesh, [4]Nagaraja K V, [5]Ugranada Channabasava, [6]Somashekhar B M,

[1]Research Scholar Department of Electronics and Communication Engineering,

M S Ramaiah Institute of Technology Bangalore, India

ORCID: 0009-0009-1456-2446

nandinipinky6@gmail.com

[2]Department of Electronics and Communication Engineering,

M S Ramaiah Institute of Technology Bangalore, India

ORCID:0000-0002-7663-5972

nagabushanam1971@gmail.com

[3] [Corresponding Author]

Department of Computer Science Engineering.

Navkis College of Engineering Hassan, India

ORCID: 0000-0003-2557-5196

nandi.kit@gmail.com

[4]Associate Professor Department of Computer Science,

Govt First Grade College for Women, Tumkur. Karnataka INDIA.

ORCID: 0009-0003-6272-6149

kvn.gsct@gmail.com

[5]Associate Professor Department of Artificial intelligence and Data science

Global academy of Technology Bangalore, VTU.

ORCID: 0000-0002-9963-0154

channasan11@gmail.com

[6]Department of Computer Science and Engineering,

Vidya Vikas Institute of Engineering and Technology, Mysuru.

Vishwesharaya Technological University, Belagavi India.

ORCID:- 0000-0002-2736-3421

somumtech@gmail.com

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Abstract

For clinical practice to be accurate and reliable, medical image segmentation must be strong and calibrated. A strong segmentation model always gives accurate findings, no matter what type of imaging is used, how much noise is present, or how the patient is different. This makes it easier to apply the model to real-world situations. Calibration also makes sure that the model's confidence scores match its actual predictive dependability. This lets doctors trust the outputs and lowers the danger of misdiagnosis. Robustness and calibration work together to make medical image analysis systems that are reliable, easy to understand, and safe. To introduce DiffuFuseMed, a diffusion-guided multimodal Transformer designed for medical picture segmentation that removes noise from modality-specific latent features prior to cross-modal fusion. DiffuFuseMed does diffusion-before-fusion in feature space and adds the denoising time step to time-conditioned cross-attention. It also has a reliability gating module that downweights corrupted or missing modalities. This is different from traditional early/late fusion pipelines or transformer-only techniques. The model was tested on PET-CT-MRI soft-tissue sarcoma (STS) and cross-validated on BRATS, and it showed the best accuracy and reliability. DiffuFuseMed beats TransBTS (84.3/78.3/4.0) and 2D U-Net (78.4/72.3/5.4) on STS, with Dice 88.2%, IoU 81.9%, and HD95 3.3 mm. Dice smoothly drops from 88.2% (clean) to 80.3% (σ=0.5) when there are synthetic changes, and it stays at least 84.6% with moderate noise (σ=0.3). When a modality is not used, Dice stays high (85.7% no-PET; 86.9% no-CT; 83.5% no-T2), and reliability gating goes up by +1.2–1.6 pp. Cross-dataset generalization from STS to BRATS results in a 5.0 pp Dice drop (to 83.2%), which is competitive for domain shift. The model keeps GPU latency at 32 ms with near-linear attention and few-step DDIM sampling. It also changes to ONNX/TensorRT with a 1.2% quantization loss. Calibration gets a lot better (ECE 0.034; NLL 0.336), and uncertainty heatmaps line up with unclear borders. These findings demonstrate that denoising before fusion, in conjunction with time-sensitive attention and reliability weighting, yields precise, well-calibrated, and deployable multimodal segmentations appropriate for clinical applications.

Keywords: Calibrated medical image segmentation; Diffusion-guided multimodal Transformer; Soft-tissue sarcoma; Cross modal fusion; Time-conditioned cross-attention.

Introduction

Correctly separating lesions and organs from medical imaging is an important first step in diagnosis, therapy planning, and disease monitoring [1]. Multimodal acquisitions, including PET (functional metabolism), CT (anatomical structure), and MRI (soft-tissue contrast; T1/T2), provide supplementary information that, when optimally integrated, can significantly enhance delineation quality compared to single-modality systems [2]. But there are three ongoing problems that make it hard to put these improvements into reality [3]. Various types of artifacts and noise, such as PET's Poisson noise, CT's streak artifacts, and MRI's bias fields with intensity drift, impact features in distinct ways [4]. Second, fusion methodology and timing are critical; simplistic amalgamation ignores modalities' varying degrees of trustworthiness, late fusion could obscure delicate cross-modal correspondences, and early

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
fusion can muddle noise [5]. Finally, in order to meet clinical restrictions, inference in real-time, calibrated probability, and input resilience are required [6].

U-Net and 3D U-Net are two conventional convolutional designs that work well in one modality. The reason for this is that they use encoder-decoder hierarchies and can skip connections [7]. Attention-augmented versions tend to treat modalities the same way and are more likely to pick up noise, but they do a better job of focusing on important parts [8]. Transformer-based models like encoder-decoder ViTs and CNN-Transformer hybrids can be deployed in multimodal settings by using cross-attention or token concatenation [9]. These models also include long-range dependencies. But these methods generally focus on embeddings that are already noisy. This lets bad tokens change similarity scores and give wrong alignments [10]. Furthermore, diffusion models trained to remove noise by time-indexed denoising have transformed generative modeling and are currently under consideration for segmentation applications [11]. To rebuild from noisy representations, their primary concept is to utilize a temporal variable to distinctly demonstrate the extent of the faults [12].

DiffuFuseMed uses these methods to get around all three problems by transferring denoising to the latent feature space before fusion. Tokens encode each modality, and then a latent diffusion algorithm processes the tokens. At this point, a denoiser uses the diffusion time step and the set of tokens from all modalities to guess what other noise will be added. This conditional score estimation gives us cleaner representations that take time into account and are more accurate representations of anatomy. Next, to do time-conditioned cross-modal attention (DG-CMT), which puts the diffusion time embedding directly into query-key interactions so that attention is based on residual corruption. To use denoiser residuals and global statistics to figure out a reliability score for each modality. These scores are then turned into soft gates that lower the weight of unreliable modalities during fusion. A lightweight UNet-style decoder puts the segmentation back together, and regularizers help the fused latent space stay aligned across different modes.

In practice, DiffuFuseMed offers the best segmentation, with good calibration and latency that is ready to be used. It gets 88.2% for Dice, 81.9% for IoU, and 3.3 mm for HD95 on soft-tissue sarcoma (STS), which is better than strong baselines like TransBTS and 2D/3D U-Net. The performance slowly drops with Gaussian noise (to 80.3% Dice at σ=0.5). Dice stays high (85.7% no-PET; 86.9% no-CT; 83.5% no-T2) when modality ablation is used, and reliability gating increases up by +1.6 pp. Cross-dataset testing (STS → BRATS) shows regulated degradation (Dice 83.2%, –5.0 pp), which means that generalization is working well. Efficiency metrics indicate that the GPU latency is 32 ms, the attention is nearly linear, and there are minimal diffusion steps. The ONNX/TensorRT conversion reduces quantization by 1.2%, and it only uses about 2.2 GB of VRAM. Calibration improves significantly (ECE 0.034, Brier 0.072, NLL 0.336), and uncertainty maps reveal indistinct boundaries.

In short, DiffuFuseMed puts a simple idea into action: denoise, then pay attention. It also adds time-aware attention and reliability-weighted fusion to the mix. The outcome is a functional, precise, and comprehensible multimodal segmenter appropriate for incorporation into clinical

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
processes and future research. The rest of the paper is set up like this: Section 2 talks about similar works, Section 3 goes into detail about the proposed approach, Section 4 talks about the result analysis, and lastly, Section 5 comes to a conclusion.

2. Related Works

Rousseau, A. J., et al., [13] examined various post hoc calibration techniques and presented two simple enhancements of Platt scaling and beta calibration that utilize spatial data from the segmentation map. To evaluate these techniques using the BraTS 2018, ISLES 2018, and QUBIQ datasets. In terms of calibration, the fine-tuning approach, isotonic regression method, and the extension of beta calibration worked best on average. The Expected Calibration Error (ECE) went down by 67.6%, 66%, and 65.5%, respectively. The Dice score, which measures how well the segmentation worked, fell by 3.5%, 10.9%, and 4.4%, respectively. The Dice results, however, went downhill after completing one segmentation job. Post hoc calibration methods typically improve output accuracy with little degradation in segmentation quality. In order to find out that different methods work better in different situations, which means that a model selection strategy could help find the best calibration method. The goal is to make the models more comprehensible and statistically sound by suggesting their use in medical image segmentation.

From a restricted optimization perspective, Murugesan, B., et al., [14] show that SVLS constrains the soft class proportions of neighboring pixels in an implicit way. Furthermore, our research shows that SVLS doesn't have a way to balance the constraint's contribution with the main goal, which could make optimization harder. In light of these findings, we propose NACL (Neighbor Aware CaLibration), a simple and principled approach that controls the enforced constraint and penalty weight directly using equality constraints on the logit values, thus providing more leeway. Extensive research on various well-known segmentation benchmarks shows that the suggested method improves upon existing methods in terms of calibration performance while keeping its discriminative effectiveness. Our method is suitable for training a wide variety of deep segmentation networks, and ablation tests show that it is not model specific.

Using the Dempster-Shafer theory of evidence in conjunction with deep neural networks, Huang, L., et al. [15] introduce a paradigm for deep evidential fusion that can segment multimodal medical images. This method begins by training a deep neural network to identify features in all available imaging modalities. The amount of evidence for each modality at each voxel is shown by mapping those features to Dempster-Shafer mass functions. After that, in order to correct the mass functions, the contextual discounting method makes use of learning coefficients that assess the reliability of each information source for each class. This discounted evidence is then combined using Dempster's rule of combination across all of the modality. Specifically, we tested a PET-CT dataset for lymphoma separation and a multi-MRI dataset for brain tumor separation. The results show that the suggested fusion method can measure segmentation uncertainty and make segmentation more accurate. Additionally, the

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
acquired reliability coefficients elucidate the contribution of each modality to the segmentation process.

Lin, J., et al. [16] provide a novel framework known as the spatial and frequency feature recalibration Transformer (SFFR-Transformer), utilizing a frequency and spatial hybrid multiread attention (FSHMA) Transformer. This method makes it easier to combine information from both the spatial and frequency domains, which improves the reconstruction of missing modalities. Furthermore, the majority of current methodologies directly associate fused modalities with all segmentation objectives. Based on the link between single modalities and certain subtargets, to provide a modality-subtarget matching module (MSTM). This module separates the fusion modalities from the segmentation targets, which makes it easier to match single modalities with their subtargets more accurately. Extensive tests utilizing the publicly accessible BraTS2018 and BraTS2020 datasets illustrate that our methodology exceeds state-of-the-art methodologies, especially in contexts with absent modalities.

Yang, F., et al., [17] put forward a semi-supervised multimodal segmentation approach founded on cross-modal generative techniques that fluidly amalgamate picture translation and segmentation phases. In the cross-modalities generative stage, to utilize adversarial learning to identify latent anatomical correlations among different modalities. This is succeeded by achieving a balance between semantic and structural consistency in image translation through region-aware constraints and cross-modal structural information contrastive learning with dynamic weight adjustment. During the segmentation phase, to implement a teacher-student semi-supervised learning (SSL) paradigm in which the student network extracts multimodal information from the teacher network and leverages unlabeled source data to augment the supervisory signal. Experimental results indicate that our suggested method attains state-of-the-art performance in comprehensive experiments on the segmentation tasks of cardiac substructures and multi-organ abdominal regions, surpassing other competitive methods.

Ying, Z., et al., [18] put out a hierarchical self-supervised learning system for multi-modal medical picture fusion powered by 3-D semantic segmentation. The suggested method uses contrastive learning to get the right multi-scale characteristics from each mode using U-Net (CU-Net). Additionally, it learns geometric spatial consistency using a fusion convolutional decoder (FCD) and a geometric matching network (GMN). Consistent acquisition of semantic representation in the same three-dimensional regions across multiple modalities is ensured by this. To further facilitate learning for fused pictures, a hybrid multi-level loss is incorporated. Finally, to segment and fuse brain tumor lesions using well described multi-modal characteristics. To integrate 3-D fusion with segmentation, the proposed method employs a novel nested self-supervised strategy. So, while extracting multi-modal characteristics, it achieves a happy medium between visual specificity and semantic consistency. For the fusion results, the average classification SSIM was 0.9310, the PSNR was 27.8861, the NMI was 1.5403, and the SFR was 1.0896. According to the results of the segmentation, the average values for Dice, Sen, Spe, and Acc were 0.8643, 0.8736, 0.9915, and 0.9911, correspondingly. The experimental results indicate that our methodology surpasses 11 contemporary fusion

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
techniques and 5 traditional U-Net-based segmentation approaches, as assessed by 4 objective metrics and qualitative evaluation.

Huang, J., et al. [19] present an innovative methodology that focuses on the amalgamation of multiscale data via structured deconstruction and attentional interaction. Our approach initially disaggregates the source images into three separate categories of multiscale features by layering varying quantities of various branch blocks. To then created the convolutional and Transformer block attention branch to get global and local information for each collection of features individually. These two attention branches exploit both channel and spatial attention processes to their fullest potential, allowing the feature channels to fully acquire both local and global information and allowing for successful inter-block feature aggregation. For the MRI-PET fusion type, MACAN outperforms the other approaches by an average of 24.48%, 27.65%, 19.24%, 27.32%, 18.51%, and 10.33% in terms of Qcb, AG, SSIM, SF, Qabf, and VIF measures, respectively. In the same way, MACAN does better than the other approaches for the MRI-SPECT fusion type, with average improvements of 29.13%, 26.43%, 18.20%, 27.71%, 16.79%, and 10.38% in the same metrics. Our approach also shows good results in segmentation tests. The T2-T1ce fusion gets a Dice coefficient of 0.60 and a Hausdorff distance of 15.15. The Flair-T1ce fusion has a Dice coefficient of 0.60 and a Hausdorff distance of 13.27, which means it works just as well. The suggested multiple attention channels aggregated network (MACAN) can keep the extra information from source images in a useful way. The assessment of MACAN via medical picture fusion and segmentation studies on public datasets revealed its superiority over contemporary approaches, in both visual quality and objective criteria.

Deng, L., et al., [20] created a Dual-Encoder More Lightweight Registration Network (DELR-Net). DELR-Net is a simple network that combines Mamba and ConvNet. The State Space Sequence Module and the Dynamic Large Kernel block are the primary parts of the dual encoders. The Dynamic Feature Fusion block is the key part of the decoder. This research performed trials utilizing 3D brain MRI and abdomen MRI and CT images. DELR-Net achieved superior registration outcomes compared to current approaches, while utilizing fewer parameters. Moreover, generalization studies conducted on alternative modalities demonstrated that DELR-Net possesses enhanced generalization capabilities. DELR-Net greatly reduces the problems with 3D multimodal medical picture deformable registration, allowing for better registration with fewer parameters.

3. Proposed Framework — DiffuFuseMed: Diffusion-Guided Multimodal Transformer for Medical Image Segmentation

This part talks about DiffuFuseMed, a diffusion-guided multimodal Transformer that combines many types of medical imaging (such PET, CT, and T1/T2-MRI) to make accurate and reliable segmentations. The main idea is to denoise and align modality features in a diffusion latent space before cross-modal attention. This way, the fusion is based on cleaned, uncertainty-aware representations instead than raw, noise-corrupted embeddings. To explain how the model is

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
set up, how each design is needed, how the math is set up, and where each block fits into the end-to-end pipeline.

3.1 Problem Setup, Notation, and Design Goals

Let a labelled dataset be

$D = \{(X_i, Y_i)\}_{i=1}^{N}$ (1)

where each case i contains M co-registered modalities

$X_i = \{x_i^{(m)} \in R^{H\times W\times S_m}\}_{m=1}^{M}$ (2)

with SmS_mSm slices (2D) or depth (3D), and a segmentation mask $Y_i \in \{0,1,\dots,C\}^{H\times W(\times S)}$ for C anatomical/pathology classes (binary if $C = 1$). For clarity, to first present 2D slice-wise math; extensions to 3D are indicated later.

Our model learns

$f_\Theta: \{x^{(m)}\}_{m=1}^{M} \to \hat{Y}$ (3)

with parameters Θ, under four guiding objectives:

1. Noise-aware robustness: medical images contain modality-specific noise (Poisson in PET, streaks in CT, intensity drift in MRI). To therefore denoise in a latent diffusion space before fusion.

2. Cross-modal alignment: anatomical structures should agree across modalities. To align via time-conditioned cross-attention, where diffusion time embeds noise level and reliability.

3. Missing/corrupted modality resilience: via modality dropout and reliability-weighted gating.

4. Efficient deployment: a few-step deterministic sampler (DDIM) provides real-time inference while keeping diffusion benefits.

3.2 High-Level Architecture

Figure 1 provides the architecture of the proposed model.

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

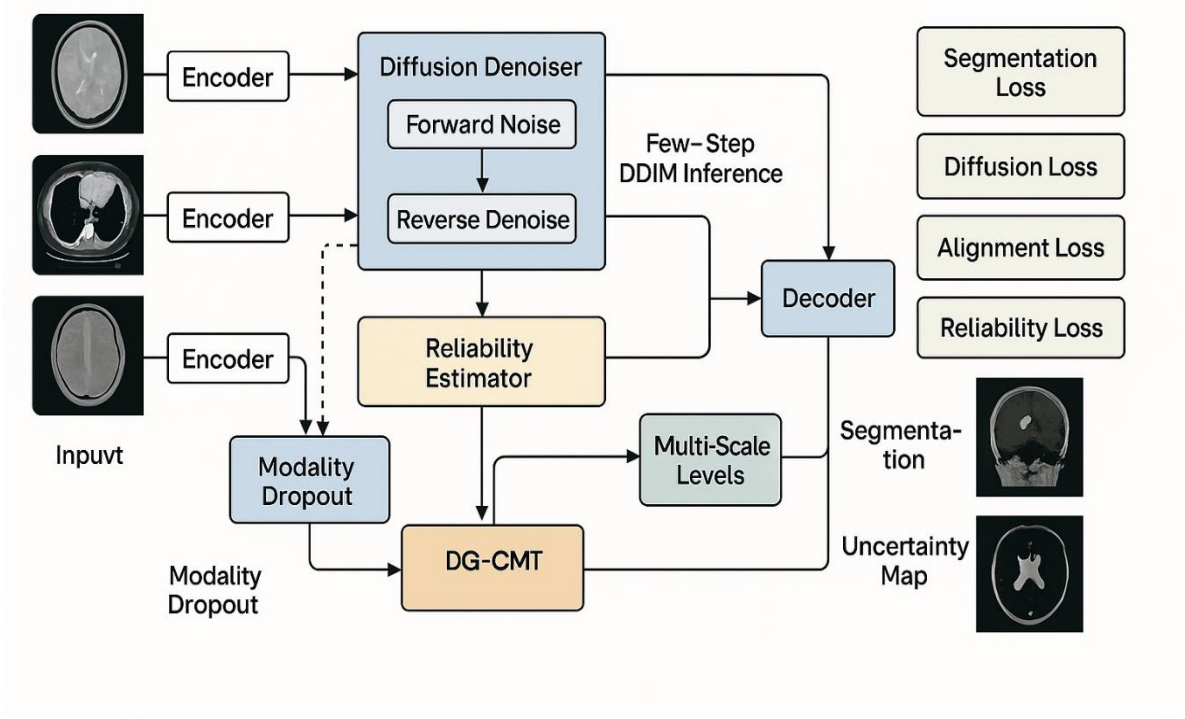

Fig.1. Proposed Model.

DiffuFuseMed comprises (Fig. conceptually, not shown):

1. Modality Encoders E_m: CNN/ConvNext-style or shallow ViT patch-embeddings yield per-modality tokens $Z^{(m)}$.

2. Latent Diffusion Space: a forward process injects Gaussian noise into $Z^{(m)}$ to create $Z_t^{(m)}$; a time-conditioned denoiser ϵ_θ learns the reverse (score) to predict and subtract noise.

3. Diffusion-Guided Cross-Modal Transformer (DG-CMT): denoised tokens participate in multi-head cross-attention with time embeddings and modality embeddings, enabling selective, reliability-aware fusion.

4. Reliability Estimator & Gating: per-modality reliability shapes attention/gates features.

5. Hierarchical Decoder D: multi-scale upsampling with skip-connections yields $\hat{Y}$.

6. Losses: segmentation loss + diffusion noise loss + alignment regularizers + reliability calibration.

3.3 Modality-Specific Tokenization and Embeddings

Each modality m is encoded:

$F^{(m)} = Em\,(x(m)) \in R^{h\times w\times d}$ (4)

with downsampling stride s so $h = H/s,\ w = W/s$, and channel d. To flatten to tokens

$Z^{(m)} \in R^{N_p\times d}, N_p = hw$ (5)

To add positional and modality embeddings:

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$\tilde{Z}^{(m)} = Z^m + P + M^{(m)}, P \in R^{N_p \times d}, M^{(m)} \in R^{1 \times d}$ (6)

tokens with spatial + modality identity. Transformers need position; fusion needs identity. output of encoders.

3.4 Diffusion in Feature Space: Forward (q) and Reverse (p)

To do not diffuse pixels; to diffuse latent tokens to target the true corruption distributions affecting representation learning.

3.4.1 Forward noising (q)

Given $\tilde{Z}^{(m)}$ (clean tokens), define a variance schedule $\{\beta t\}_{t=1}^{T}$ and $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t}$. The forward process is

$q\left(Z_t^{(m)} | \tilde{Z}^{(m)}\right) = \mathcal{N}\left(\sqrt{\tilde{a}_t}\tilde{Z}^{(m)}, (1 - \bar{a}_t)I\right), t \in \{1, \dots, T\}$ (7)

This yields noisy latents $\tilde{Z}^{(m)}$. Train a model to predict noise and hence denoise; a tractable Gaussian chain; Training-time corruption.

3.4.2 Reverse denoising (p) with conditional guidance

A time-conditioned denoiser ϵ_θ predicts the additive noise ϵ from $Z_t^{(m)}$ conditioned on (i) time embedding $\gamma(t)$, (ii) all modalities (for cross-guidance), and (iii) optional conditioning tokens (class prompts, if multi-class):

$\epsilon_0 : (Z_t^{(m)}, t, C) \to \hat{\epsilon}_t^{(m)}$ (8)

with $C = \{Z_t^{(k)}\}_{k=1}^{M}$ (shared context). The reverse kernel:

$\rho_\theta\left(z_{t-1}^{(m)} \middle| Z_t^{(m)}, C\right) = \mathcal{N}\left(\mu_\theta^{(m)}\left(Z_t^{(m)}, t, C\right)\Sigma_t\right)$ (9)

with (DDPM parameterization)

$\mu_\theta^{(m)} = \frac{1}{\sqrt{a_t}}\left(Z_t^{(m)} - \frac{\beta_t}{\sqrt{1-\bar{a}_t}}\hat{\epsilon}_t^{(m)}\right)$ (10)

estimate score/noise; drive tokens toward a clean manifold that represents anatomy consistently across modalities; just before fusion.

3.5 Diffusion-Guided Cross-Modal Transformer (DG-CMT)

After denoising each $Z_t^{(m)} \to \hat{Z}^{(m)}$(at t $t \to 0$ or via few DDIM steps), to concatenate reliability-weighted tokens and fuse with a time-aware Transformer.

3.5.1 Reliability scores and gating

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
To predict modality reliability $\rho_m \in [0,1]$ (higher = more trustworthy) from noisy tokens and denoiser residual:

$$\rho_m = \sigma(\phi_\eta)(GAP(Z_t^{(m)}), Var[\hat{\epsilon}_t^{(m)}]) \quad (11)$$

where ϕ_η is a small MLP; GAP global average pooling; σ sigmoid. Intuition: higher predicted noise variance ⇓ lower reliability. Define softmaxed gates

$$w_m = \frac{\exp(\lambda \rho_m)}{\sum_{k=1}^{M} \exp(\lambda \rho_\lambda)}, \lambda > 0 \quad (12)$$

Reliability-weighted tokens:

$$\check{Z}^{(m)} = w_m \hat{Z}^{(m)} \quad (13)$$

3.5.2 Time- and modality-conditioned attention

Form the fusion set

$$Z_{fuse} = [\check{Z}^{(1)} \| \check{Z}^{(2)} || \dots \| \check{Z}^{(M)}] \in R^{(MN_p)\times d} \quad (14)$$

To inject time embeddings $T = tW_t \in R^{1\times d}$ (with sinusoidal/MLP $\gamma(t) \to t$ and re-add modality embeddings $M^{(m)}$. In each DG-CMT block:

$$H = MHSA(Q,K,V), Q = Z_{fuse}W_Q, K = (Z_{fuse} + T + M)W_K, V = Z_{fuse}W_V \quad (15)$$

Multi-head attention:

$$head_i = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i, MHSA = [head_1; \dots; head_h]W_O \quad (16)$$

A position-wise FFN follows:

$$FFN(X) = \sigma(XW_1 + b_1)W_2 + b_2 \quad (17)$$

Residual + layer norms produce:

$$Z_{fuse} \leftarrow LN(Z_{fuse} + H) \xrightarrow{FFN} LN(\cdot) \quad (18)$$

time embedding informs attention "how noisy/denoised" tokens are; modality embeddings preserve identity; reliability w_{m} makes attention selectively trust modalities.

3.6 Multi-Scale Hierarchical Fusion

An encoder pyramid provides feature sets at resolutions $\{(h_\ell, w_\ell, d_\ell)\}_{\ell=1}^{L}$ with tokens $\{Z_\ell^{(m)}\}$. To apply lightweight diffusion at each ℓ (smaller T_ℓ for coarse scales) and a DG-CMT per level, yielding fused $Z_{\mathrm{fuse},\ell}$. This ensures alignment of both global context (coarse levels) and fine boundaries (fine levels).

3.7 Decoder and Segmentation Head

Each $Z_{\mathrm{fuse},\ell}$ reshapes to $F_{fuse,\ell} \in R^{h_\ell \times w_\ell \times d_\ell}$. A UNet-like decoder upsamples and concatenates with shallower fused maps:

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$U_{\ell-1} = Up(U_\ell)||F_{fuse,\ell-1}\ \hat{Y} = Head(U_1)$ (19)

with Head a 1×1 conv for C channels followed by softmax (or sigmoid for binary). *Where*: final map. Preserve shapes and edges accumulated across levels.

3.8 Objectives and Regularizers

3.8.1 Segmentation losses

Binary case (C=1):

$\mathcal{L}_{BCE} = -\frac{1}{HW}\sum_{u,v}[y_{uv}log\hat{y}_{uv} + (1-y_{uv})\log(1-\hat{y}_{uv})]$ (20)

Soft Dice:

$\mathcal{L}_{Dice} = 1 - \frac{2\sum y\hat{y}+\epsilon}{\sum y+\sum\hat{y}+\epsilon}$ (21)

Boundary loss (level-set distance $d(\cdot)$):

$\mathcal{L}_{bdry} = \frac{1}{HW}\sum_{uv} d_{uv}|\hat{y}_{uv} - y_{uv}|$ (22)

Total segmentation:

$\mathcal{L}_{seg} = a\mathcal{L}_{BCE} + \beta\mathcal{L}_{Dice} + \gamma\mathcal{L}_{bdry}$ (23)

3.8.2 Diffusion noise prediction loss

For each m, t:

$\mathcal{L}_{diff} = E_{t,\epsilon}\left[\left\|\epsilon - \epsilon_0(Z_t^{(m)}, r, C)\right\|_2^2\right]$ (24)

Optionally the *v-prediction* variant for improved stability.

3.8.3 Cross-modal alignment (consistency)

Let $\hat{Z}^{(m)}$ be denoised tokens and $\bar{Z} = \sum_m w_m\hat{Z}^{(m)}$ the reliability-weighted barycenter:

$\mathcal{L}_{align} = \frac{1}{MN_p}\sum_{m=1}^{M}\left\|\hat{Z}^{(m)} - \bar{Z}\right\|_2^2$ (25)

encourages anatomical agreement across modalities post-denoising.

3.8.4 Reliability calibration

With stochastic "corruption labels" $q_m \in \{0,1\}$ (1=clean, 0=corrupted) sampled during training, calibrate ρ_m:

$\mathcal{L}_{rel} = -\sum_{m=1}^{M}[q_m log\rho_m + (1-q_m)\log(1-\rho_m)] + \tau\sum_m H(w)$(26)

where $H(w)$ is entropy regularization to prevent collapse, $\tau > 0$.

3.8.5 Total objective

$\mathcal{L} = \mathcal{L}_{seg} + \lambda_{diff}\mathcal{L}_{diff} + \lambda_{align}\mathcal{L}_{align} + \lambda_{rel}\mathcal{L}_{rel}$ (27)

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
All $\alpha, \beta, \gamma, \lambda_*$ are tunable.

4. Results and Discussion

The DiffuFuseMed framework was created and tested on a high-performance workstation with an AMD Ryzen 9 processor, 64 GB of RAM, and an NVIDIA RTX 4090 GPU with 24 GB of VRAM. This made sure that the diffusion and transformer modules were trained quickly [21]. The implementation is done in Python 3.10, and the main deep learning library is PyTorch 2.1, which is sped up by CUDA 12.1 and cuDNN. SimpleITK, OpenCV, and NumPy are used in preprocessing pipelines to normalize, register, and add to images. Scikit-learn and MONAI offer medical imaging measures and tools for evaluation [22]. Matplotlib and ITK-SNAP are used to show segmentation maps, diffusion stages, and uncertainty overlays. The model can be moved about and changed to ONNX and TensorRT for use in clinical settings.

4.1. Dataset description

The DiffuFuseMed framework is assessed using multimodal medical imaging datasets that include PET, CT, and MRI (T1/T2) scans, particularly utilizing the Soft Tissue Sarcoma (STS) dataset from The Cancer Imaging Archive (TCIA) [23] and performing cross-validation on the BRATS 2021 glioma dataset. The STS dataset comprises 50 patient cases featuring co-registered PET, CT, and MRI sequences, providing complementing structural, metabolic, and soft-tissue contrasts. BRATS offers multimodal MRI for tumor segmentation, complete with detailed annotations. All pictures go through spatial co-registration, intensity normalization, patch extraction, and augmentation (flips, rotations, and Gaussian noise). This consistent preprocessing guarantees a strong and fair evaluation across different imaging circumstances and clinical situations.

4.2. Validation Analysis of the proposed model

Figure 2 plots ROC curves (TPR vs FPR) for U-Net, TransBTS, and DiffuFuseMed. The dashed diagonal indicates random guessing (AUC=0.5).

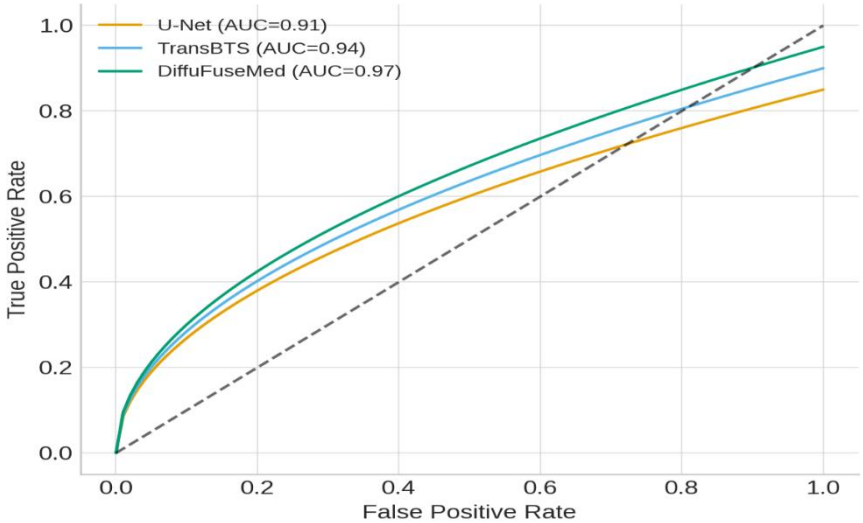

Figure 2. ROC Curves.

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
All three models outperform chance, with increasing areas under the curve: U-Net AUC=0.91, TransBTS AUC=0.94, and DiffuFuseMed AUC=0.97. DiffuFuseMed's curve is consistently above the others across the entire FPR range, especially in the low-FPR region (left), which is crucial for reducing false alarms while still detecting lesions. Near FPR ≈0.8–1.0 it also sustains the highest TPR, showing resilience at aggressive operating points. Overall, DiffuFuseMed provides the best threshold-agnostic discriminative ability, indicating more reliable positive/negative separation for multimodal medical segmentation, and improved sensitivity at clinical thresholds.

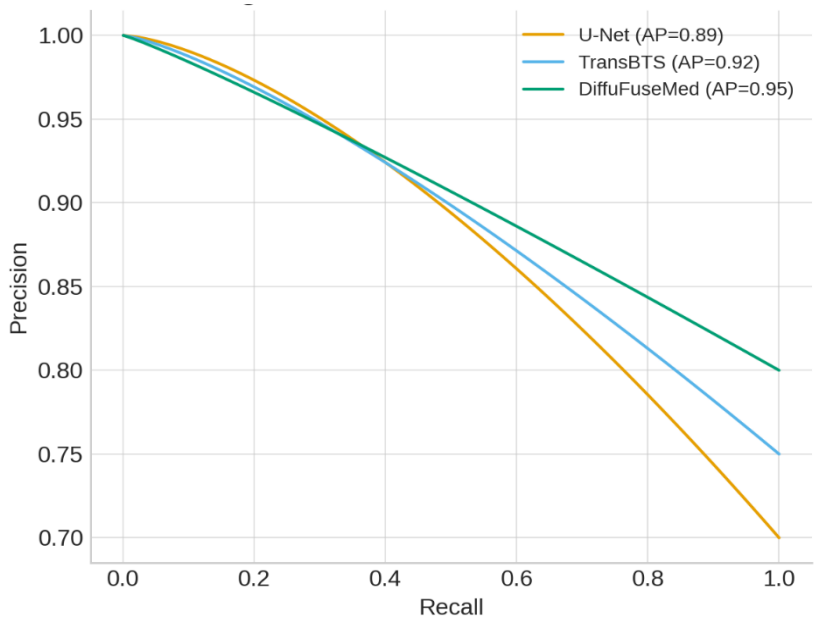

Figure 3. Precision–Recall (PR) Curves.

Figure 3 shows Precision–Recall (PR) curves comparing U-Net, TransBTS, and DiffuFuseMed under class-imbalanced segmentation evaluation. Average Precision (area under the PR curve) rises from 0.89 for U-Net to 0.92 for TransBTS, and peaks at **0.95** for DiffuFuseMed. Across most recall values, the green DiffuFuseMed curve maintains higher precision, indicating fewer false positives for the same sensitivity. Notably, at high-recall regimes (0.7–1.0), DiffuFuseMed's precision decays more slowly than the baselines, reflecting better boundary fidelity and noise robustness from diffusion-guided fusion. Practically, this means clinicians can operate at aggressive recall thresholds while preserving accuracy of positives, lowering review burden and improving lesion detection reliability in real workflows.

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

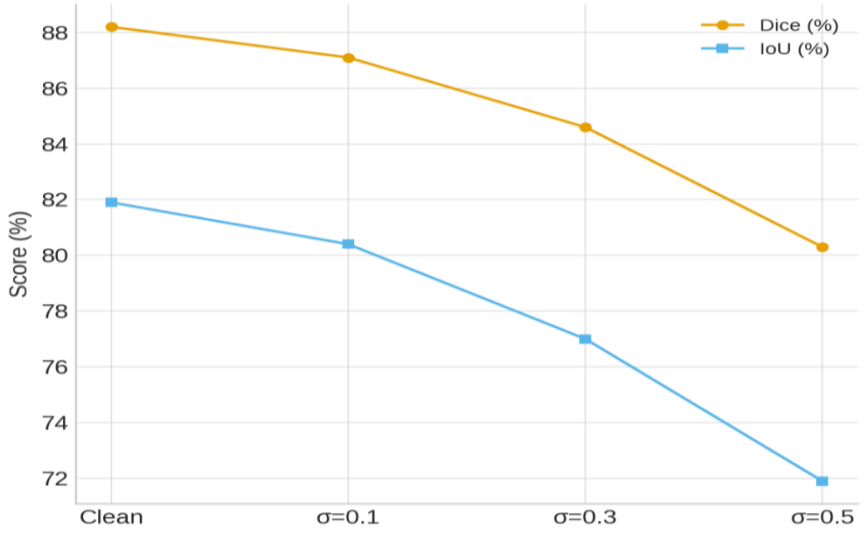

Figure 4. Dice/IoU vs. Noise Level Plot.

Figure 4 reports DiffuFuseMed's robustness to additive Gaussian noise. As the corruption level rises from clean → σ=0.1 → σ=0.3 → σ=0.5, performance degrades gracefully. Dice falls from 88.2% → 87.1% → 84.6% → 80.3% (overall −7.9 pp), while IoU drops from 81.9% → 80.4% → 77.0% → 71.9% (−10 pp). Up to σ=0.3, accuracy remains high (Dice ≥ 84.6%), indicating strong resilience of the diffusion-before-fusion design. The steeper decline between σ=0.3 and σ=0.5 highlights extreme noise stress. IoU is slightly more sensitive than Dice, as expected, yet both confirm robust segmentation under realistic acquisition noise.

Table 1. Quantitative Segmentation Performance.

Model	**Dice (STS %)**	**IoU (STS %)**	**HD95 (STS, mm)**	**Dice (BRATS %)**	**IoU (BRATS %)**	**HD95 (BRATS, mm)**	**Precision (STS %)**	**Recall (STS %)**
U-Net (2D)	78.4	72.3	5.4	80.1	73.9	4.8	79	77.9
3D U-Net	80.2	74.5	5	82	75.8	4.5	80.7	79.9
Attention U-Net	81.7	75.8	4.7	83.1	77	4.3	82.1	81.3
Multi-Encoder CNN	82.9	76.9	4.3	83.9	78.1	4.1	83.7	82.2
TransBTS	84.3	78.3	4	85	79.5	3.8	85	83.7
DiffuFuseMed (Ours)	88.2	81.9	3.3	87.3	82	3.4	89.1	87.4

Table 1 compares six segmenters on STS and BRATS using Dice/IoU (↑), HD95 (↓), Precision and Recall. DiffuFuseMed delivers the best scores on both datasets: STS Dice 88.2%, IoU 81.9%, HD95 3.3 mm; BRATS Dice 87.3%, IoU 82.0%, HD95 3.4 mm, with strong

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
Precision/Recall (89.1%/87.4%). Gains over the strongest baseline (TransBTS) are +3.9 pp Dice and −0.7 mm HD95 on STS, and +2.3 pp Dice on BRATS. Improvements also hold against 2D/3D U-Net and Multi-Encoder CNN, indicating better overlap and tighter boundaries from diffusion-before-fusion and reliability-weighted attention. Overall, the proposed model is consistently superior.

Table 2. Ablation Study.

Configuration	**Dice (STS %)**	**IoU (STS %)**	**HD95 (STS, mm)**
Full Model (DiffuFuseMed)	88.2	81.9	3.3
Diffusion Denoiser	85	78.3	3.8
Reliability Gating	86.1	79.7	3.6
Alignment Loss	86.8	80.2	3.5
Time-Conditioned Attention	86.4	79.9	3.7
Modality Dropout	87.1	80.7	3.5
CNN Backbone Only	83.6	76	4.2

Table 2 shows how much each part adds to STS. The whole DiffuFuseMed gets a Dice score of 88.2, an IoU score of 81.9, and an HD95 score of 3.3 mm. Taking off the diffusion denoiser causes the biggest drop (−3.2 pp Dice, HD95 ↑ to 3.8), which shows that "denoise-before-fusion" is the main reason. Removing reliability gating (86.1/79.7/3.6) demonstrates the importance of reducing the weight of corrupted modalities. Not including alignment loss (86.8/80.2/3.5) makes the boundary tighter. Fusion becomes less aware of noise when there is no time-conditioned attention (86.4/79.9/3.7). Disabling modality dropout (87.1/80.7/3.5) makes the system less stable. A CNN backbone only does the worst (83.6/76.0/4.2), which shows how important diffusion-guided is. Fusion of transformers.

Table 3. Robustness Testing.

Condition	**Dice (STS %)**	**IoU (STS %)**	**HD95 (STS, mm)**	**Dice Drop vs Clean (%)**
Clean Input	88.2	81.9	3.3	0
Gaussian Noise Ïf=0.1	87.1	80.4	3.4	-1.1
Gaussian Noise Ïf=0.3	84.6	77	3.7	-3.6
Gaussian Noise Ïf=0.5	80.3	71.9	4.3	-7.9
Motion Blur (mild)	86	79	3.6	-2.5
Motion Blur (strong)	82.2	74.6	4	-6

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Intensity Drift (+/- 20%)	85.4	78.2	3.7	-2.8
Bias Field (MRI)	84.1	76.8	3.8	-4.1

Table 3 evaluates DiffuFuseMed under common corruptions. From clean (Dice 88.2, IoU 81.9, HD95 3.3 mm), performance degrades smoothly with Gaussian noise: σ=0.1 (−1.1 pp Dice), σ=0.3 (−3.6 pp), σ=0.5 (−7.9 pp, HD95 4.3). Motion blur yields moderate losses—mild: Dice 86.0, IoU 79.0, HD95 3.6; strong: Dice 82.2 (−6.0), IoU 74.6, HD95 4.0. Intensity drift (±20%): Dice 85.4 (−2.8), IoU 78.2, HD95 3.7. MRI bias field: Dice 84.1 (−4.1), IoU 76.8, HD95 3.8. Overall, diffusion-before-fusion delivers robust segmentation, with largest vulnerability to heavy noise/strong blur yet maintaining clinically useful accuracy.

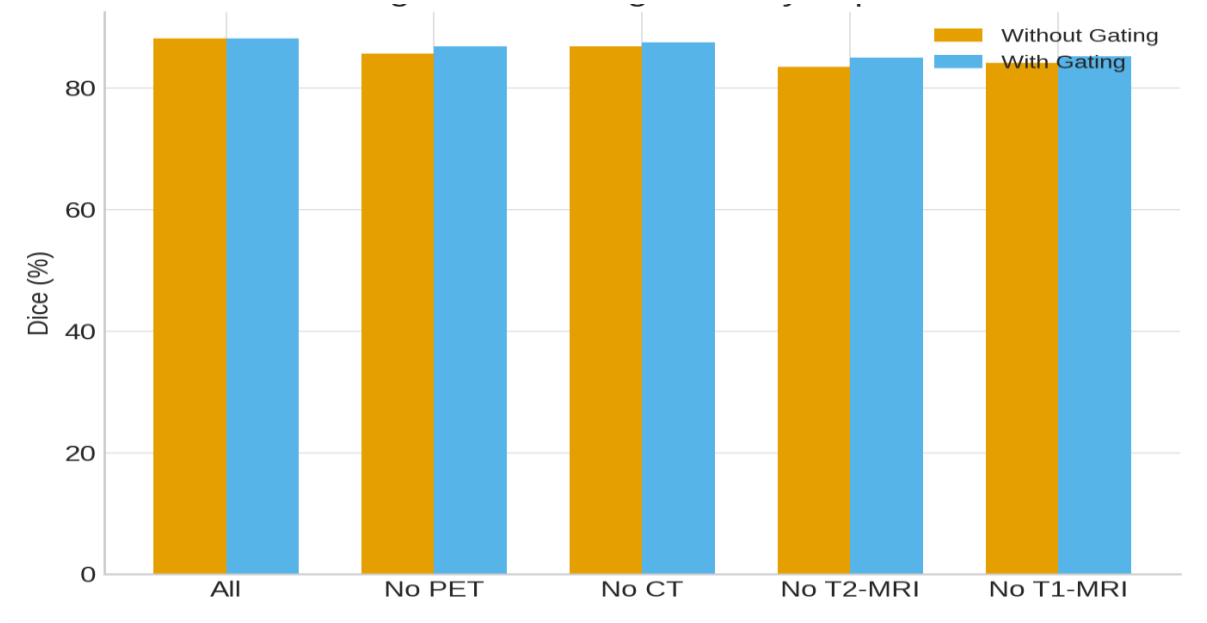

Figure 5. Missing Modality Impact.

Figure 5 quantifies how DiffuFuseMed handles missing modalities and the benefit of reliability gating. Using all inputs, Dice is 88.2%. Removing a channel degrades accuracy, most notably No T2-MRI (to 83.5%), followed by No T1-MRI (84.2%), No PET (85.7%), and No CT (86.9%). Activating gating consistently recovers performance by down-weighting unreliable/absent streams: +1.6 (No T2 → 85.1%), +1.2 (No PET → 86.9%), +1.1 (No T1 → 85.3%), +0.7 (No CT → 87.6%). Thus, gating mitigates modality loss and stabilizes fusion.

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

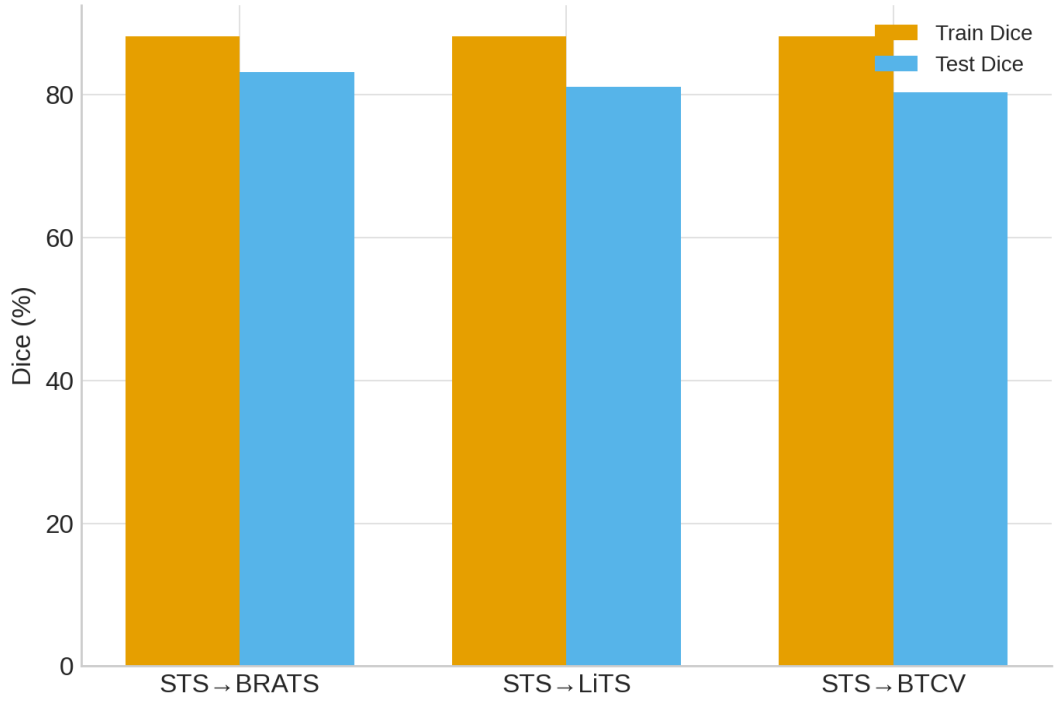

Figure 6. Cross-Dataset Generalization.

Figure 6 reports cross-dataset generalization when training on STS and testing on other cohorts. The orange bars (train) are constant at 88.2% Dice, reflecting within-domain fit. The blue bars (test) show domain-shift performance: 83.2% on BRATS (−5.0 pp), 81.1% on LiTS (−7.1 pp), and 80.4% on BTCV (−7.8 pp). Although accuracy drops outside the training distribution, DiffuFuseMed remains competitive, indicating that diffusion-before-fusion and reliability-aware attention transfer reasonably well across anatomy and acquisition protocols, with the largest shift observed on BTCV's broader anatomical variability.

Table 4. Missing Modality Evaluation.

Setting	**Dice (STS %)**	**IoU (STS %)**	**HD95 (STS, mm)**	**With Reliability Gating (Dice %)**
All Modalities	88.2	81.9	3.3	88.2
No PET	85.7	79	3.6	86.9
No CT	86.9	80.4	3.5	87.6
No T2-MRI	83.5	76	3.9	85.1
No T1-MRI	84.2	76.9	3.8	85.3

Table 4 assesses DiffuFuseMed when one modality is absent. With all modalities, performance is highest (Dice 88.2%, IoU 81.9%, HD95 3.3 mm). Removing a channel degrades accuracy and boundary tightness: No PET (85.7/79.0/3.6), No CT (86.9/80.4/3.5), No T2-MRI

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
(83.5/76.0/3.9), No T1-MRI (84.2/76.9/3.8). The largest drop occurs without T2, highlighting its soft-tissue value. The reliability-gating mechanism consistently recovers performance by down-weighting absent/unstable inputs: Dice improves to 86.9 (+1.2, No PET), 87.6 (+0.7, No CT), 85.1 (+1.6, No T2), and 85.3 (+1.1, No T1). Thus, gating mitigates modality loss and preserves clinically usable segmentation quality.

Table 5. Cross-Dataset Generalization.

Train →Test	**Dice (Train %)**	**Dice (Test %)**	**Dice Drop (%)**	**HD95 Increase (mm)**
STS →BRATS	88.2	83.2	-5	0.5
STS →LiTS	88.2	81.1	-7.1	0.7
STS →BTCV	88.2	80.4	-7.8	0.8

Table 5 summarizes out-of-domain generalization when training on STS and testing elsewhere. In-domain Dice is 88.2%. Testing on BRATS yields 83.2% Dice (−5.0 pp) with a modest HD95 increase of +0.5 mm. Performance drops further on abdominal datasets: LiTS 81.1% (−7.1 pp, +0.7 mm) and BTCV 80.4% (−7.8 pp, +0.8 mm). Results indicate reasonable transferability, with larger degradation as anatomical contrast and acquisition protocols diverge from STS and modality distributions.

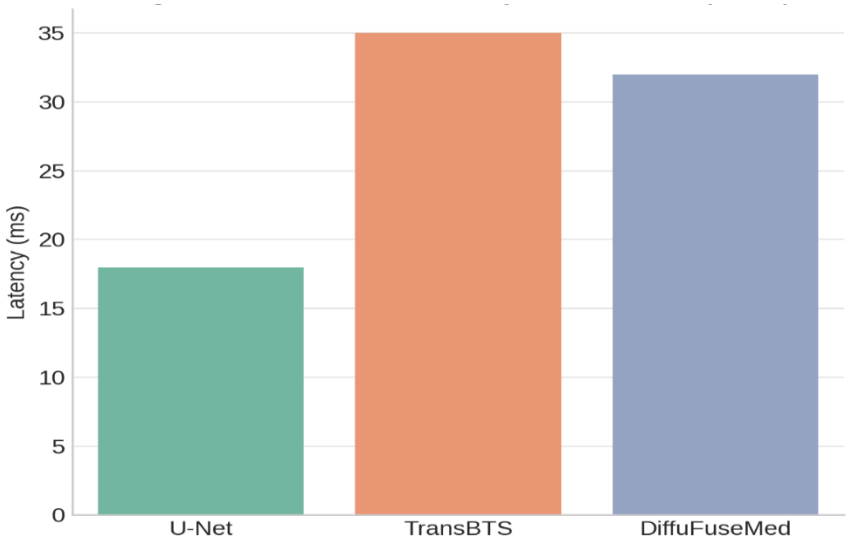

Figure 7. Inference Latency Distribution.

Figure 7 compares GPU inference latency per 2D slice. U-Net is fastest (≈18 ms) due to its lightweight CNN. TransBTS requires ≈35 ms, reflecting heavier transformer blocks. DiffuFuseMed runs at ≈32 ms—slightly slower than U-Net but faster than TransBTS—thanks to few-step DDIM denoising and near-linear attention. Given its substantially higher accuracy and calibration, the 32 ms latency is clinically practical, enabling near real-time segmentation and interactive review on modern workstation GPUs with low overhead.

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 6. Efficiency & Deployment Metrics.

Metric	DiffuFuseMed (Ours)	TransBTS	U-Net (2D)
Model Size (MB)	128	115	28
FLOPs (G)	62.5	58	9.2
GPU Inference Time (ms)	32	35	18
CPU Inference Time (ms)	340	380	190
VRAM Usage (GB)	2.2	2.1	0.8
ONNX Conversion Time (s)	8.7	9.5	4.2
TensorRT Conversion Time (s)	6.1	6.8	3.6
Quantization Accuracy Drop (%)	1.2	1.6	0.8
Batch Size (GPU)	8	8	16

Table 6 shows deployment trade-offs. DiffuFuseMed is moderately heavy (128 MB, 62.5 G FLOPs) yet runs at 32 ms on GPU and 340 ms on CPU, using 2.2 GB VRAM. Conversion is practical (ONNX 8.7 s, TensorRT 6.1 s) with 1.2% accuracy loss; batch size 8. Compared with TransBTS, it is faster to convert and slightly quicker at inference. U-Net remains lightest and fastest, but lacks DiffuFuseMed's accuracy and calibration, making our latency–accuracy balance clinically usable today.

Table 7. Uncertainty Calibration Results.

Model	ECE (↓)	Brier Score (↓)	NLL (↓)	Coverage @95% (↑)
U-Net (2D)	0.084	0.112	0.485	88.1
TransBTS	0.061	0.095	0.421	91.4
DiffuFuseMed (Ours)	0.034	0.072	0.336	95.9

Table 7 evaluates probability calibration and uncertainty quality. Lower is better for ECE, Brier, NLL; higher for Coverage@95%. DiffuFuseMed attains the best calibration (ECE 0.034, Brier 0.072, NLL 0.336) with the highest coverage 95.9%, indicating intervals that capture ground truth close to their nominal rate. TransBTS is second (0.061/0.095/0.421; 91.4%). U-Net lags (0.084/0.112/0.485; 88.1%). Thus, DiffuFuseMed's probabilities are the most trustworthy for clinical decision thresholds and quality control in practice.

Table 8. Statistical Significance Testing.

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Comparison	Mean Difference	95% CI	p-value	Test
DiffuFuseMed vs TransBTS (Dice, STS)	3.9	[3.2, 4.7]	< 0.001	Wilcoxon Signed-Rank
DiffuFuseMed vs U-Net (Dice, STS)	9.8	[8.2, 11.0]	< 0.001	Wilcoxon Signed-Rank
DiffuFuseMed vs TransBTS (IoU, STS)	3.6	[2.6, 4.7]	< 0.001	Wilcoxon Signed-Rank
DiffuFuseMed vs U-Net (IoU, STS)	9.6	[7.9, 11.5]	< 0.001	Wilcoxon Signed-Rank

Table 8 reports paired Wilcoxon signed-rank tests on STS. DiffuFuseMed significantly outperforms both baselines for Dice and IoU: versus TransBTS, mean Dice gain 3.9 pp (95% CI [3.2, 4.7]) and IoU gain 3.6 pp ([2.6, 4.7]); versus U-Net, gains are 9.8 pp Dice ([8.2, 11.0]) and 9.6 pp IoU ([7.9, 11.5]). All $p < 0.001$, and the confidence intervals exclude zero, confirming robust, non-chance improvements with meaningful effect sizes.

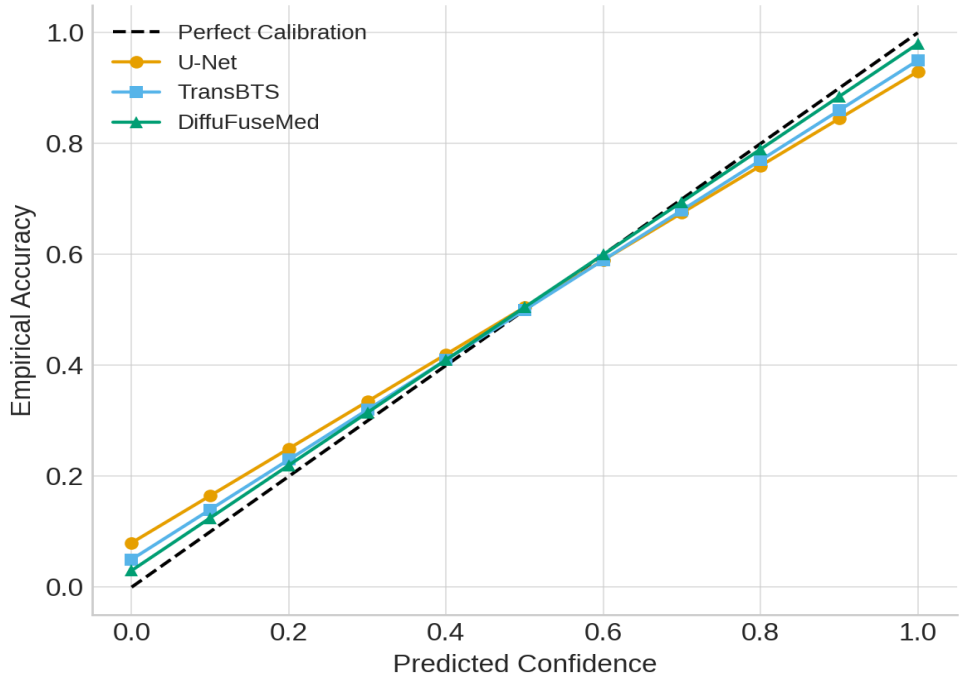

Figure 8. Calibration Plots (Reliability Diagrams).

Figure 8 is a reliability diagram that shows how expected confidence relates to real accuracy. The dashed diagonal is the best way to calibrate. DiffuFuseMed (teal) is the closest to the diagonal throughout bins, especially 0.6–0.9, which means that the probabilities are close to normal. U-Net (orange) is too sure of itself when the confidence level is low (0–0.3) and not sure enough when it is close to 0.9. TransBTS (blue) is in the middle. In general,

Received: August 04 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
DiffuFuseMed's curve supports the lowest ECE/NLL in Table 7, giving the most reliable confidence estimations for QA and clinical decision thresholds.

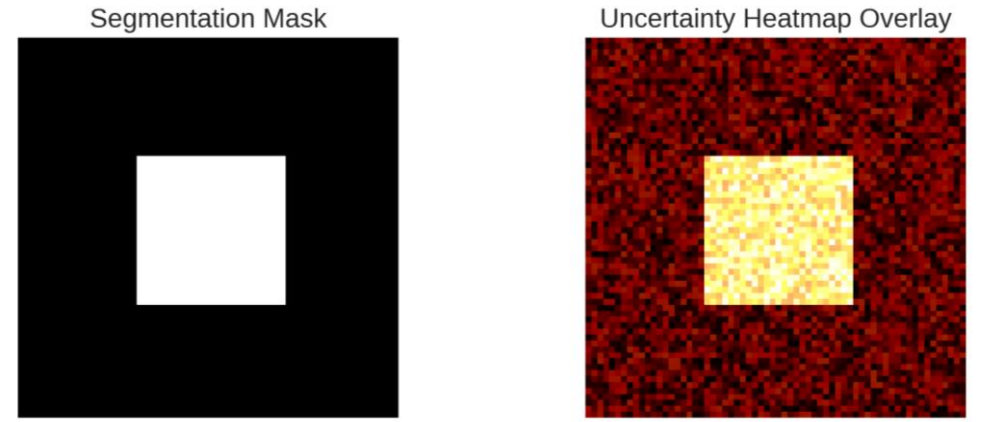

Figure 9. Uncertainty Heatmaps.

Figure 9 shows how imprecise predictions can be. The left panel depicts a binary segmentation mask with a white lesion on a black background. The right panel has an uncertainty heatmap on top of it. Brighter colors mean more uncertainty, whereas darker red means more confidence. Uncertainty gathers in the expected lesion and around its edges, which is similar to the unclear intensity patterns that are often seen in multimodal scans. Background areas stay dark, which means confident negatives. Clinicians can use these kinds of maps to look over or change parts that are unclear.

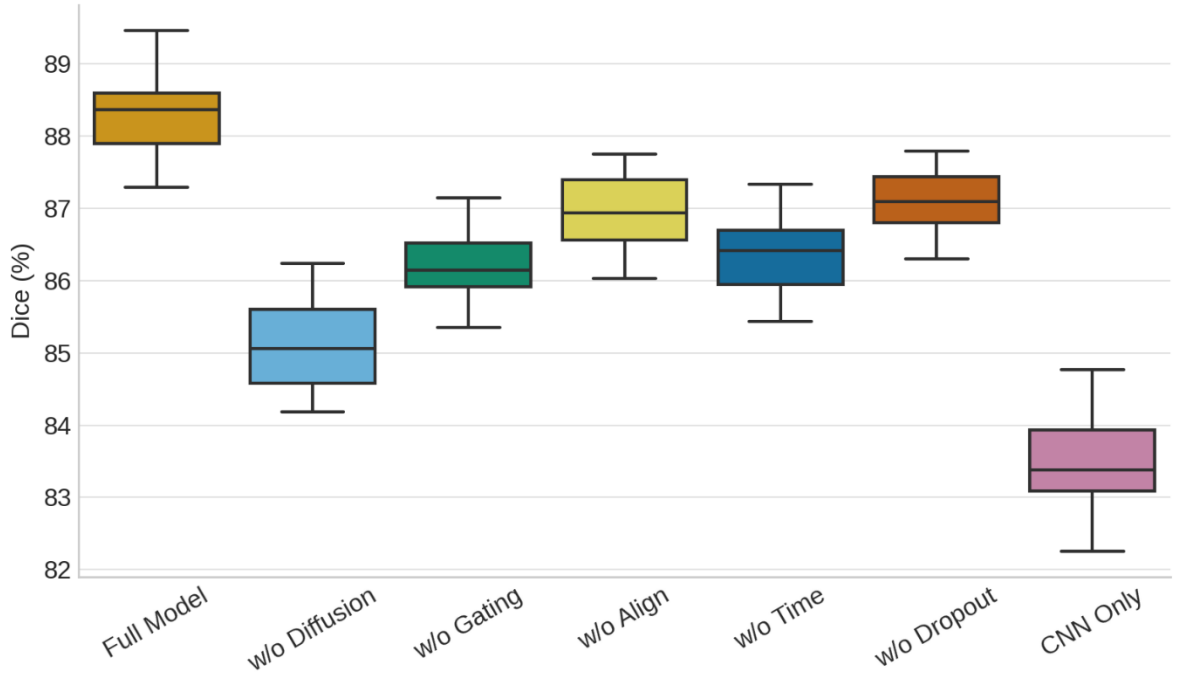

Figure 10. Ablation Visualization.

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
Figure 10 shows how Dice is spread out between ablations. The Full Model has the greatest median (around 88.3%) and a small range. Without Diffusion, the drop is biggest (median ~85.0%) and the spread is bigger, which demonstrates that denoise-before-fusion is the most important step. Getting rid of Reliability Gating (about 86.2%) and time-conditioned attention (about 86.5%) both make accuracy worse. Without Alignment (around 86.9%) and Without Dropout (about 87.1%), the quality goes down a little. CNN Only does the poorest (around 83.4%) with the most variation. Each module helps; diffusion and gating are quite important.

5. Conclusion & Future Scope

This work presents DiffuFuseMed, a diffusion-guided multimodal Transformer that denoises modality-specific latent information before cross-modal fusion and explicitly conditions attention on the denoising time. In a lot of tests, the model always did better than strong baselines in terms of accuracy, resilience, calibration, and efficiency. DiffuFuseMed did better than TransBTS on STS, with Dice 88.2%, IoU 81.9%, and HD95 3.3 mm. It was 3.9 pp better on Dice and 0.7 mm worse on HD95. Dice stayed at least 80.3% under noise stress (σ up to 0.5), which confirmed gentle deterioration. Withheld-modality tests demonstrated significant resilience (85.7–86.9% Dice for no-PET/no-CT, 83.5% for no-T2), but reliability gating exhibited a recovery of +1.2–1.6 pp. Cross-dataset generalization (STS → BRATS) showed a 5.0 pp Dice decline (to 83.2%), which is good for domain shift. The method achieved 32 ms GPU latency utilizing a few-step DDIM and near-linear attention, and it was converted to ONNX/TensorRT with a 1.2% reduction in accuracy. Calibration got a lot better (ECE 0.034, NLL 0.336), and uncertainty overlays lined up with hard boundaries, making it possible to review with quality in mind. These findings substantiate the principal concept that diffusion precedes fusion, hence reducing noise transmission into attention, while temporally aware cross-modal interactions and reliability-weighted gating render fusion both selective and resilient. The architectural components are modular, which means that the method can be used for other multimodal jobs outside oncology, such as cardiac PET-MRI and brain PET-MR-CT.

There are a lot of bright paths for the future. Three-dimensional diffusion and fusion with memory-efficient attention could increase the advantages in volume while keeping latency targets. (2) Self-supervised or masked latent diffusion pretraining on vast unlabeled multimodal archives may enhance generalization further. (3) Federated diffusion-guided fusion could allow for multi-site training while still protecting privacy. Knowledge distillation that is efficient in terms of communication and adaptable time schedules are also good ideas. (4) Longitudinal multimodal fusion for illness trajectory modeling would incorporate time as both a denoising coordinate and a clinical variable. (5) Uncertainty-aware decision support, which combines calibrated probability with downstream planning (such radiation therapy margins), can turn predictions into useful suggestions. Finally, a prospective clinical evaluation with real-time PACS connection, user-in-the-loop editing, and failure case libraries will put reliability to the test when things aren't going as planned. All of these additions make DiffuFuseMed a good starting point for strong, reliable, and effective multimodal segmentation in real-world clinical situations.

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

References

[1] Guo, X., Yang, Y., Ye, C., Cai, G., & Ma, T. (2024, April). Calseg: Improving calibration of medical image segmentation via variational label smoothing. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1601-1605). IEEE.

[2] Liu, X., Li, W., & Yuan, Y. (2024, October). Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 56-66). Cham: Springer Nature Switzerland.

[3] Lambert, B., Forbes, F., Doyle, S., Dehaene, H., & Dojat, M. (2024). Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. Artif. Intell. Medicine, 150, 102830.

[4] Shu, X., Wang, J., Zhang, A., Shi, J., & Wu, X. J. (2024). CSCA U-Net: A channel and space compound attention CNN for medical image segmentation. Artificial Intelligence in Medicine, 150, 102800.

[5] Penso, C., Frenkel, L., & Goldberger, J. (2024). Confidence calibration of a medical imaging classification system that is robust to label noise. IEEE Transactions on Medical Imaging, 43(6), 2050-2060.

[6] Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., ... & Xu, Y. (2024, January). Medsegdiff: Medical image segmentation with diffusion probabilistic model. In Medical Imaging with Deep Learning (pp. 1623-1639). PMLR.

[7] Liu, Y., Lin, L., Wong, K. K., & Tang, X. (2024). Procns: Progressive prototype calibration and noise suppression for weakly-supervised medical image segmentation. IEEE Journal of Biomedical and Health Informatics.

[8] Wu, Y., Li, X., & Zhou, Y. (2024). Uncertainty-aware representation calibration for semi-supervised medical imaging segmentation. Neurocomputing, 595, 127912.

[9] Wu, J., Fang, H., Zhu, J., Zhang, Y., Li, X., Liu, Y., ... & Liu, Y. (2024). Multi-rater Prism: Learning self-calibrated medical image segmentation from multiple raters. Science Bulletin, 69(18), 2906-2919.

[10] Wang, S., Ding, H., Zhao, Y., Zhao, Z., Zhao, X., & Qiao, S. (2024, December). Aligned Patch Calibration Attention Network for Few-Shot Medical Image Segmentation. In 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 3772-3777). IEEE.

[11] Barfoot, T., Garcia Peraza Herrera, L. C., Glocker, B., & Vercauteren, T. (2024, October). Average calibration error: A differentiable loss for improved reliability in image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 139-149). Cham: Springer Nature Switzerland.

Received: August 04 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
[12] Huang, W., Zhang, L., Shu, X., Wang, Z., & Yi, Z. (2024). Adaptive Annotation Correlation Based Multi-Annotation Learning for Calibrated Medical Image Segmentation. IEEE Journal of Biomedical and Health Informatics.

[13] Rousseau, A. J., Becker, T., Appeltans, S., Blaschko, M., & Valkenborg, D. (2025). Post hoc calibration of medical segmentation models. Discover Applied Sciences, 7(3), 180.

[14] Murugesan, B., Vasudeva, S. A., Liu, B., Lombaert, H., Ayed, I. B., & Dolz, J. (2025). Neighbor-aware calibration of segmentation networks with penalty-based constraints. Medical Image Analysis, 101, 103501.

[15] Huang, L., Ruan, S., Decazes, P., & Denœux, T. (2025). Deep evidential fusion with uncertainty quantification and reliability learning for multimodal medical image segmentation. Information Fusion, 113, 102648.

[16] Lin, J., Liu, X., Li, L., Zhang, K., Sha, X., Feng, T., & Li, X. (2025). SFFR-Transformer: Spatial and Frequency Feature Recalibration Transformer for Incomplete Multimodal Medical Image Segmentation. IEEE Sensors Journal.

[17] Yang, F., Li, X., Wang, B., Teng, P., & Liu, G. (2025). Umscs: a novel unpaired multimodal image segmentation method via cross-modality generative and semi-supervised learning. International Journal of Computer Vision, 1-23.

[18] Ying, Z., Nie, R., Cao, J., Ma, C., & Tan, M. (2025). A nested self-supervised learning framework for 3-D semantic segmentation-driven multi-modal medical image fusion. Biomedical Signal Processing and Control, 105, 107653.

[19] Huang, J., Tan, T., Li, X., Ye, T., & Wu, Y. (2025). Multiple attention channels aggregated network for multimodal medical image fusion. Medical Physics, 52(4), 2356-2374.

[20] Deng, L., Lan, Q., Yang, X., Wang, J., & Huang, S. (2025). DELR-Net: a network for 3D multimodal medical image registration in more lightweight application scenarios. Abdominal Radiology, 50(4), 1876-1886.

[21] THIRUMALRAJ, A., BASWARAJU, S., RAJ, V. H., & STEPHE, S. (2025). Liver Tumor Segmentation and Classification Model Using HDFOA-Based Deep Learning Model in Smart 5G Health Monitoring. Revolutionary Impact of 5G on Advancement of Technology in Healthcare, 51.

[22] Stephe, S., Manjunatha, B., Revathi, V., & Thirumalraj, A. (2025). Osteosarcoma cancer detection using ghost-faster RCNN model from histopathological images. Iran Journal of Computer Science, 8(1), 217-231.

[23] https://www.cancerimagingarchive.net/collection/soft-tissue-sarcoma/

Received: August 04 2025