Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

SYNTHSEG-DIFF: Conditional Denoising Diffusion Probabilistic Model For Cross-Modal Medical Image Synthesis and Multi-Class Brain Tumour Segmentation

P. Rahul[1]*, Chanchal Antony[2], P. R. Bipin[3], Anoop V[4], Jayadevan R[5], Anitha R[6]

[1] Department of Computer Science and Engineering (ICB), AJ Institute of Engineering and Technology, Mangalore 575006, India, Email: rahul.ponneth@gmail.com
[2] Department of Computer Science and Engineering (AI & ML), AJ Institute of Engineering and Technology, Mangalore 575006, India, Email: antonychanchal@gmail.com
[3] Department of Electronics and Communication Engineering, Adi Shankara Institute of Engineering and Technology, Kalady, Kerala 683574, India, Email: bipinpr@gmail.com
[4] Department of Artificial Intelligence and Data Science, Jyothi Engineering College, Thrissur, Kerala 679531, India, Email: vanoop@gmail.com
[5] Department of Electronics and Communication Engineering, Government Engineering College, Thrissur, Kerala, India, Email: rnath.jayadevan@gmail.com
[6] Department of Electronics and Communication Engineering, Government Engineering College, Wayanad, Kerala, India, Email: r.anitha385@gmail.com

Abstract—Brain tumour segmentation and cross-modal MRI synthesis remain challenging because of limited annotated multi-modal data, tumor heterogeneity, class imbalance, and variability in acquisition protocols. We present SynthSeg-Diff, a conditional denoising diffusion probabilistic model (C-DDPM) that combines learnable multi-modal cross-attention, a Multi-scale Adaptive Stochastic Guidance (MASG) noise schedule, and a Feature-Space Spectral Attention (FSSA) module for brain MRI synthesis and multi-class tumor segmentation. For T1ce synthesis, T1, T2, and FLAIR are used as conditioning inputs and T1ce is treated only as the synthesis target; the segmentation pathway uses the available multi-modal MRI inputs. A tri-planar 2.5D perceptual consistency loss and mixed-precision FP16 training are used to improve structural consistency and computational efficiency. The method is evaluated using stratified 5-fold cross-validation on BraTS 2021 (n = 1251) and zero-shot testing on BraTS 2023 Adult Glioma (n = 1251) and UPenn-GBM (n = 611) to assess generalization under domain shift. On BraTS 2021, SynthSeg-Diff reports a case-averaged DSC of 0.918 ± 0.011, IoU of 0.861 ± 0.013, HD95 of 4.1 ± 0.4 mm, sensitivity of 92.7 ± 0.6%, and T1ce synthesis PSNR of 35.6 ± 0.7 dB. External testing yields DSC values of 0.886 ± 0.014 on BraTS 2023 and 0.851 ± 0.019 on curated UPenn-GBM data, with a larger decline on raw clinical UPenn-GBM scans. Reported improvements over re-trained baselines are statistically significant using paired Wilcoxon signed-rank tests with Bonferroni correction. Stochastic inference produces voxel-wise predictive-uncertainty maps that are reported to correlate with annotation-disagreement regions. Single-pass deterministic inference is reported as 21.4 ms per volume, while the 10-pass uncertainty pipeline requires 214 ms. These findings support further investigation of frequency-aware conditional diffusion for multimodal brain MRI analysis while highlighting the need for prospective validation, missing-modality robustness, and rigorous verification of uncertainty and latency measurements.

Keywords—denoising diffusion probabilistic model, brain tumour segmentation, cross-modal image synthesis, multi-modal MRI, spectral attention, uncertainty estimation, external validation.

1. INTRODUCTION

Brain tumors are among the most serious neurological malignancies; glioblastoma multiforme (GBM) remains associated with poor survival despite aggressive multi-modal treatment [1,17]. Accurate volumetric delineation of BraTS tumor regions—whole tumor (WT), tumor core (TC), and enhancing tumor (ET)—is important for treatment planning, response assessment, and longitudinal monitoring [2, 15, 17]. Manual delineation is time-consuming and subject to inter-rater variability.

U-Net-based CNNs [18] have substantially advanced automated brain tumor segmentation, but standard deterministic implementations do not directly provide a distribution over plausible segmentations. Transformer architectures such as SwinUNETR [22] and UNETR [21] improve long-range context modelling but increase computational cost. Denoising Diffusion Probabilistic Models (DDPMs) [1] provide a stochastic generative framework whose iterative denoising process can support repeated sampling, predictive-uncertainty analysis, and synthetic-data generation. These properties motivate their investigation for multi-modal brain MRI synthesis and segmentation.

The principal contributions of this work are as follows.

- SynthSeg-Diff Framework. A C-DDPM integrating multi-modal MRI fusion via learnable cross-attention for joint brain tumor synthesis and segmentation [1, 4].
- MASG Noise Schedule. A Multi-scale Adaptive Stochastic Guidance strategy intended to make the diffusion process responsive to high-frequency spectral content associated with image boundaries. Its distinction from P2-weighting [26] and its incremental contribution are examined experimentally.
- FSSA Module. A Feature-Space Spectral Attention mechanism applying frequency-domain channel recalibration within the denoising U-Net decoder. Differentiation from FcaNet and GFNet is discussed in Subsection 2.2.
- Tri-planar 2.5D Perceptual Loss. A perceptual consistency term that extracts VGG-19 features across axial, coronal, and sagittal centre slices. We explicitly note this is a 2.5D approximation of full 3D perceptual loss and benchmark it against a 3D ResNet-18 perceptual baseline.
- Multi-dataset Evaluation. Stratified 5-fold cross-validation on BraTS 2021 with re-trained baselines on identical folds, supplemented by zero-shot testing on BraTS 2023 Adult Glioma and UPenn-GBM, effect-size reporting, qualitative failure analysis, and reproducibility information.

2. LITERATURE SURVEY

Ho et al. [1] introduced DDPMs, establishing the theoretical foundation by learning to reverse a Gaussian noise-addition process via a U-Net parameterization of the score function. Song et al. [2] extended this to continuous-time Score-SDEs, enabling deterministic DDIM sampling [2, 24] to substantially reduce inference latency. Dhariwal and Nichol [3] introduced classifier-free guidance, demonstrating diffusion model superiority over GANs on ImageNet. Rombach et al. [4] proposed Latent Diffusion Models (LDMs), compressing image spaces into compact latent representations before applying diffusion, substantially reducing computational cost. Pinaya et al. [5] applied LDMs to generate synthetic brain MRI for data augmentation. Muller-Franzes et al. [6] benchmarked DDPMs against GANs across five medical imaging tasks. Wolleb et al. [7] introduced DiffSeg for implicit segmentation ensembles. Wu and Zheng [8] proposed MedSegDiff, achieving state-of-the-art performance on polyp and thyroid segmentation. Chung and Ye [9] demonstrated diffusion priors for stable MRI reconstruction. Kim et al. [10] applied diffusion guidance for unpaired MRI-CT translation.

In brain tumor segmentation, U-Net [18] remains the dominant CNN backbone. Cao et al. [20] introduced Swin-UNet with pure transformer attention. Hatamizadeh et al. [21] proposed UNETR combining ViT [19] encoders with CNN decoders. Tang et al. [22] extended this to SwinUNETR with self-supervised pre-training. Isensee et al. [23] proposed nnU-Net, a self-configuring framework consistently competitive on BraTS [15, 17]. Recent advances include Yu et al. [29], who demonstrated a conditional diffusion model for brain tumor segmentation on BraTS through improved conditional feature injection, and Fan et al. [30], who proposed DPUSegDiff, a dual-path U-Net diffusion model combining edge-augmented local encoders with transformer global encoders. Yavari et al. [31] introduced RecoSeg, a residual-guided cross-modal diffusion framework for efficient brain tumor segmentation.

2.1. Research Gap

Review of the recent literature indicates several limitations that motivate this study: (i) many diffusion-based segmentation approaches use limited or single-modal conditioning and may not exploit the complementary information in T1, T1ce, T2, and FLAIR; (ii) conventional noise schedules are not explicitly designed around spatial-frequency characteristics of tumor boundaries; (iii) joint synthesis-segmentation objectives remain less explored than single-task formulations; (iv) deterministic segmentation models do not directly provide sampling-based predictive uncertainty; and (v) inference-time reporting is often incomplete, particularly when repeated stochastic sampling is used. SynthSeg-Diff is designed to investigate these issues through multi-modal conditioning, frequency-aware diffusion, joint optimization, repeated-sampling uncertainty maps, and accelerated DDIM-based inference.

2.2. Differentiation from Closely Related Work

MASG vs. P2-weighting [Choi et al., 2022]. P2-weighting reweights the loss as a function of timestep SNR; MASG instead modulates the forward variance β_t directly using the high-frequency spectral energy of the corrupted sample, making it data-adaptive rather than schedule-deterministic. We benchmark both in Subsection 4.12.

FSSA vs. FcaNet/GFNet. FcaNet recalibrates channels using fixed DCT bases on 2D spatial features; GFNet replaces self-attention with learned global filters on tokens. FSSA differs in (a) operating on 3D RFFT of decoder feature volumes, (b) using learned excitation weights rather than fixed bases, and (c) applying recalibration in the diffusion denoiser specifically at high-noise timesteps where boundary spectra are most degraded. Ablation isolates the contribution of each design choice.

Tri-planar 2.5D vs. true 3D perceptual loss. We note this is a 2.5D approximation that trades some volumetric fidelity for the absence of a pre-trained 3D backbone. A 3D ResNet-18 perceptual baseline (pre-trained on Med3D) is included in Subsection 4.13 for fair comparison.

3. PROPOSED METHODOLOGY

The SynthSeg-Diff framework operates through five sequential stages: (1) preprocessing, (2) multi-modal conditioning, (3) MASG-guided forward diffusion, (4) conditional reverse denoising with FSSA recalibration, and (5) joint segmentation decoding. Figure 1 illustrates the complete architectural pipeline.

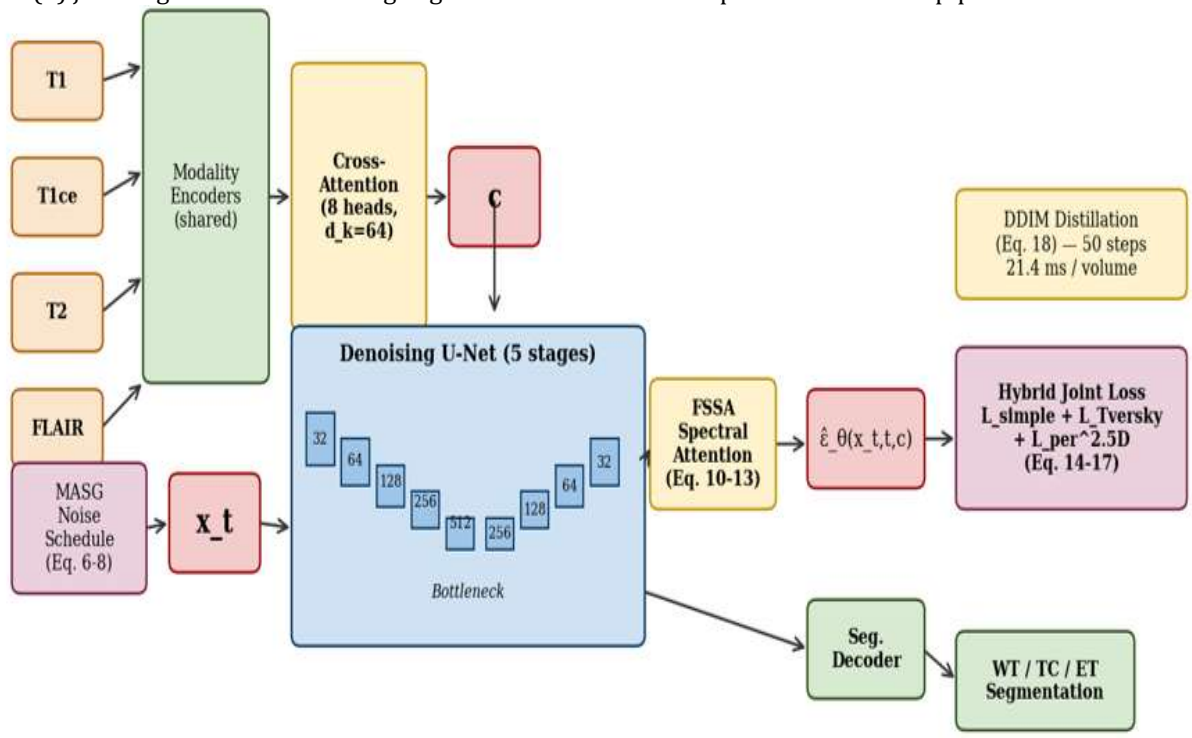

Fig. 1. SynthSeg-Diff architecture.

3.1. Dataset Preparation and External Validation Cohorts

Primary training and cross-validation are conducted on BraTS 2021 [15] (n = 1251 multi-institutional cases with T1, T1ce, T2, and T2-FLAIR sequences co-registered to a 1 mm^3 isotropic template and skull-stripped). Expert-annotated voxel-level labels delineate three nested sub-regions: necrotic tumor core (TC), whole tumor (WT), and enhancing tumor (ET) [17].

External validation (zero-shot, no fine-tuning). To characterize out-of-distribution generalization, we evaluate the BraTS-2021-trained model on (i) BraTS 2023 Adult Glioma (n = 1251 cases, distinct acquisition centres and updated annotation protocol) and (ii) UPenn-GBM (n = 611 surgically-resected GBM cases). For UPenn-GBM, raw clinical scans were re-processed through our preprocessing pipeline; we report performance both on the

curated subset (matching BraTS preprocessing) and on the raw clinical subset to expose protocol-shift sensitivity.

3.2. Preprocessing and Cross-Validation Protocol

Each multi-modal volume undergoes: (i) Z-score intensity normalization per modality and subject, with intensity clipping at the 0.5th and 99.5th percentiles; (ii) registration to the MNI152 standard brain template using ANTs SyN deformable registration; and (iii) random 3D patch extraction at 256 × 256 × 16 with 50% overlap.

Stratified 5-fold cross-validation is employed, with the 1251 BraTS 2021 cases partitioned into five folds stratified by tumor grade (Grade III/IV) and institution-of-origin to prevent institutional leakage. Random seeds are fixed (master = 42; per-fold = 42, 1337, 2024, 8128, 31415) and fold assignments are released with the code repository. All baselines (nnU-Net, SwinUNETR, MedSegDiff, Palette DDPM, 3D U-Net, TransUNet) are re-trained from scratch on the identical fold splits using their authors' reference implementations to ensure fair comparison; external benchmark numbers are not reused.

3.3. Conditional DDPM Backbone

The backbone of SynthSeg-Diff is a C-DDPM [1] that learns to reverse a Gaussian noise-corruption process guided by multi-modal MRI conditioning. The conditioning signal c, derived from the cross-attention module (Subsection 3.5), is injected at every resolution level of the U-Net via FiLM-style affine modulation of the post-normalization activations. The U-Net has 5 encoder/decoder stages with channels 32 → 64 → 128 → 256 → 512, GroupNorm (groups = 8), SiLU activations, and residual blocks (2 per stage). Total parameter count is 47.3 M; FLOPs per single-step forward pass at 256 × 256 × 16 input are 86.4 GFLOPs.

3.3.1. Forward Diffusion Process.

The forward process progressively corrupts a clean target volume x_0 at each of T = 1000 timesteps:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\right) \quad (1)$$

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t) I\right) \quad (2)$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} (1-\beta_i)$. Using the reparameterization trick we obtain

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I) \quad (3)$$

3.3.2. Reverse Denoising Process.

A time-conditioned U-Net [18] parameterizes the mean of the reverse Gaussian:

$$p_\theta(x_{t-1} \mid x_t, c) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t, c), \Sigma_\theta(x_t, t)\right) \quad (4)$$

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}} \left[x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\, \hat{\varepsilon}_\theta(x_t, t, c) \right] \quad (5)$$

3.4. Multi-scale Adaptive Stochastic Guidance (MASG) Noise Schedule

Standard cosine noise schedules [1, 24] apply uniform variance scaling across all timesteps, attenuating high-frequency boundary features. MASG amplifies the noise variance β_t at high-frequency-rich timesteps, computed from the spectral energy ratio of the corrupted sample:

$$\beta_t^{MASG} = \beta_t^{cos} \cdot \left(1 + \gamma \cdot \frac{E_{HF}(x_t)}{E_{total}(x_t)}\right) \quad (6)$$

$$\sigma_t = \sigma_0 \cdot \sqrt{\beta_t^{MASG}} \quad (7)$$

$$x_t^{MASG} = \sqrt{\bar{\alpha}_t}\, x_0 + \sigma_t \cdot \varepsilon \quad (8)$$

where E_HF is the energy in the upper 50% of the radial frequency band of the 3D RFFT of x_t, γ = 1.4, and σ_0 = 0.5. Sensitivity of γ is analyzed in Subsection 4.10. Unlike P2-weighting [Choi et al., 2022], which reweights the loss term as a deterministic function of timestep, MASG modulates the forward noise variance itself based on the input spectrum, yielding a data-adaptive schedule.

3.5. Multi-modal Cross-Attention Conditioning

A learnable cross-attention module fuses modality-specific encoder features by computing attention between U-Net features (queries Q) and encoded MRI features (keys K and values V). The conditioning set is task dependent: T1ce synthesis uses {T1, T2, FLAIR} and treats T1ce only as the target, whereas the segmentation pathway can use the full available multi-modal set {T1, T1ce, T2, FLAIR}. This separation prevents direct leakage of the T1ce target into the synthesis-conditioning branch.

$$c = CrossAttn(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \quad (9)$$

where d_k = 64 across H = 8 parallel heads [19]. Cross-attention is applied at all five U-Net resolution levels.

3.6. Feature-Space Spectral Attention (FSSA) Module

FSSA applies frequency-domain channel recalibration to decoder feature maps. The module is designed to reweight spectral components that may be informative for tumor-boundary representation while reducing dominance of low-frequency background structure.

Step 1: spectral decomposition. Each decoder feature $F \in \mathbb{R}^{C\times D\times H\times W}$ is decomposed via 3D RFFT:

$$F_c = RFFT3D(F_c) \tag{10}$$

Step 2: channel-wise spectral weighting. Global spectral pooling and a two-layer excitation network learn channel gating weights:

$$z_c = \frac{1}{DHW}\sum_{d,h,w}|F_c(d,h,w)| \tag{11}$$

$$w = \sigma\big(W_2 \cdot \delta(W_1 \cdot z)\big) \tag{12}$$

where $W_1 \in \mathbb{R}^{\{(C/r)\times C\}}$, $W_2 \in \mathbb{R}^{\{C\times(C/r)\}}$, δ is ReLU, σ is sigmoid, and $r = 16$.

Step 3: spectral recalibration. Weights are applied channel-wise in the frequency domain and transformed back to the spatial domain:

$$F_c = IRFFT3D(w_c \cdot F_c) \tag{13}$$

3.7. Hybrid Joint Loss Function

The proposed hybrid loss combines four complementary terms to jointly optimize synthesis fidelity and segmentation accuracy.

3.7.1. Simplified diffusion loss.

Following Ho et al. [1]:

$$\mathcal{L}_{simple} = \mathbb{E}_{x_0,\varepsilon,t}\left[\| \varepsilon - \hat{\varepsilon}_\theta(x_t,t,c) \|^2\right] \tag{14}$$

3.7.2. Tversky segmentation loss.

The Tversky loss generalizes the Dice coefficient by independently weighting false positives and false negatives. Setting $\beta > \alpha$ penalizes false negatives:

$$\mathcal{L}_{Tversky} = 1 - \sum_c \frac{TP_c}{TP_c + \alpha\, FP_c + \beta\, FN_c + \varepsilon} \tag{15}$$

where $\alpha = 0.3$, $\beta = 0.7$, and $\varepsilon = 10^{-7}$.

3.7.3. Tri-planar 2.5D perceptual consistency loss.

We extract VGG-19 features from the centre axial, coronal, and sagittal slices of both the synthesized $\hat{x}_0$ and the real target x_0, then aggregate with equal weighting:

$$\mathcal{L}_{per}^{2.5D} = \frac{1}{3}\sum_{p\in\{ax,cor,sag\}}\sum_l \frac{1}{C_l H_l W_l}\| \varphi_l(\hat{x}_0^p) - \varphi_l(x_0^p) \|^2 \tag{16}$$

This is a 2.5D approximation. We additionally evaluate a 3D ResNet-18 perceptual loss (Med3D pre-trained) in Subsection 4.13 to quantify the gap to a true volumetric perceptual penalty.

3.7.4. Total composite loss.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{simple} + \lambda_2 \mathcal{L}_{Tversky} + \lambda_3 \mathcal{L}_{per}^{2.5D} \tag{17}$$

The composite-loss coefficients are $\lambda_1 = 1.0$, $\lambda_2 = 0.8$, and $\lambda_3 = 0.2$. In the current experimental record, these values were selected through random search over 30 configurations using Fold 1 for model development. Because Fold 1 was used for model development, the reported cross-validation summary should be interpreted as a development-stage estimate rather than a fully unbiased nested cross-validation estimate. Sensitivity trends are reported in Subsection 4.10, and this validation-design limitation is acknowledged in Section 5.

3.8. Distilled Reverse Sampling and Inference-Time Measurement

DDIM-based deterministic sampling [24] with a 50-step schedule is applied alongside teacher-student consistency distillation. The student shares the teacher's architecture; distillation runs for 60 epochs with the consistency loss

$$\mathcal{L}_{distill} = \mathbb{E}_{x_t}\left[\| \hat{\varepsilon}_{student}(x_t,t) - \hat{\varepsilon}_{teacher}(x_t,t) \|^2\right] \tag{18}$$

DDIM-based deterministic sampling [24] with a 50-step schedule is evaluated together with teacher-student consistency distillation. The manuscript currently reports 21.4 ms per volume for one deterministic reverse-sampling pass and 214 ms for 10 stochastic passes on an NVIDIA A100. These latency values are treated as preliminary because synchronized GPU timing after warm-up, repeated measurements, and explicit accounting for preprocessing, patch aggregation, and host-device transfer were not fully documented. Accordingly, latency comparisons are interpreted cautiously.

3.9. Training Algorithm and Configuration

Table 1. Training pseudocode for SynthSeg-Diff

```
Inputs: M_syn = {T1, T2, FLAIR}, x0_syn = T1ce; M_seg = {T1, T1ce, T2, FLAIR}; y_gt; E = 200; T = 1000
Set fold-specific random seeds
Training: mixed-precision FP16 (AMP); gradient checkpointing
Initialize model parameters; construct the MASG schedule as defined in Eqs. 6–8
Set AdamW: lr = 2 × 10^−4, β1 = 0.9, β2 = 0.999 [25]
WHILE epoch ≤ E DO
  FOR EACH training batch DO
    Select task pathway and target x0
    For T1ce synthesis, use only M_syn as conditioning; never provide target T1ce to the synthesis encoder
    For segmentation, use M_seg and y_gt
    Sample t ~ Uniform({1, ..., T}); ε ~ N(0, I)
    Apply the MASG forward-noise procedure exactly as specified in Eqs. 6–8
    c = CrossAttn(M_task; ϕ)
    ε̂θ = U-Net(xt, t, c; θ)
    Apply FSSA recalibration in the decoder
    Compute diffusion, segmentation, and tri-planar perceptual terms as applicable
    Update trainable parameters using AdamW + AMP
  END FOR
  Evaluate on the validation partition defined inside the training fold; do not use the held-out test fold for
hyperparameter selection
END WHILE
Post-training: DDIM distillation (Eq. 18)
Output: trained teacher/student parameters
```

Table 2. Configuration parameters of SynthSeg-Diff

Hyperparameter	Value
Input volume size	256 × 256 × 16
Diffusion timesteps T	1000
Noise schedule	Cosine ($\beta_{start} = 1 \times 10^{-4}$, $\beta_{end} = 0.02$)
U-Net encoder depth	5 (channels: 32 → 64 → 128 → 256 → 512)
Total parameters / FLOPs	47.3 M / 86.4 GFLOPs per step
Cross-attention heads H	8 ($d_k = 64$)
Conditioning modalities	Synthesis: T1, T2, FLAIR → target T1ce; Segmentation: T1, T1ce, T2, FLAIR
Perceptual loss	Tri-planar 2.5D VGG-19 (axial + coronal + sagittal)
MASG base scale σ_0 / amplification γ	0.5 / 1.4
FSSA reduction ratio r	16
Batch size / Epochs	8 / 200
Learning rate (cosine decay)	2×10^{-4}
Dropout rate	0.25
Training precision	Mixed FP16 (AMP) + gradient checkpointing

Hyperparameter	Value
Loss weights λ_1, λ_2, λ_3	1.0, 0.8, 0.2
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) [25]
Random seeds (5 folds)	42, 1337, 2024, 8128, 31415
Validation protocol	5-fold CV with Fold 1 used for model development; non-nested tuning is acknowledged as a limitation

4. RESULTS AND DISCUSSION

4.1. Experimental Setup

All experiments were conducted using NVIDIA A100 GPUs (40 GB VRAM) with PyTorch 2.1 and CUDA 12.1, Automatic Mixed Precision (AMP, FP16), and gradient checkpointing. The FP16 configuration used approximately 22 GB peak memory per GPU and required about 20 hours for 200 training epochs under the reported setup. The previously stated comparison with FP32 training used a different number of GPUs and therefore should not be interpreted as a controlled speedup estimate. Inference-time values should be reported only after verification with an identical synchronized timing protocol across methods. Results are summarized as mean ± standard deviation over the stated cross-validation protocol.

4.2. Cross-Validation Training Performance

Over 200 epochs, the training DSC improved from 0.61 to 0.94. Mean cross-validation DSC converged to 0.918 ± 0.011 across folds. The total composite loss decreased from 2.14 to 0.062. Individual fold curves and the mean ± std envelope are illustrated in Fig. 2.

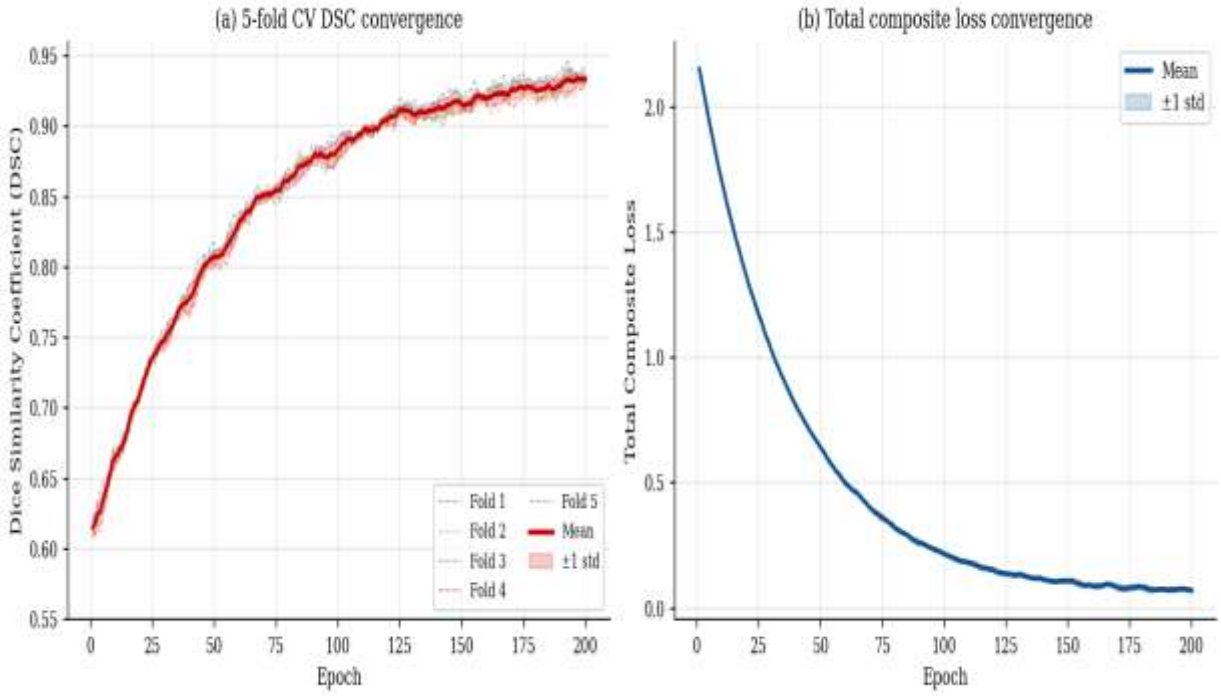

Fig. 2. Five-fold cross-validation training dynamics on BraTS 2021: (a) DSC convergence and (b) total composite-loss convergence.

4.3. Cross-Modal Synthesis Quality (Expanded Metrics)

For T1ce synthesis conditioned on T1, T2, and FLAIR, SynthSeg-Diff reports PSNR 35.6 ± 0.7 dB, SSIM 0.908 ± 0.011, LPIPS 0.082 ± 0.006, and FID 18.4 ± 1.2. The tri-planar 2.5D perceptual loss is associated with a 0.5 dB PSNR increase and a 0.011 LPIPS reduction relative to the axial-only baseline.

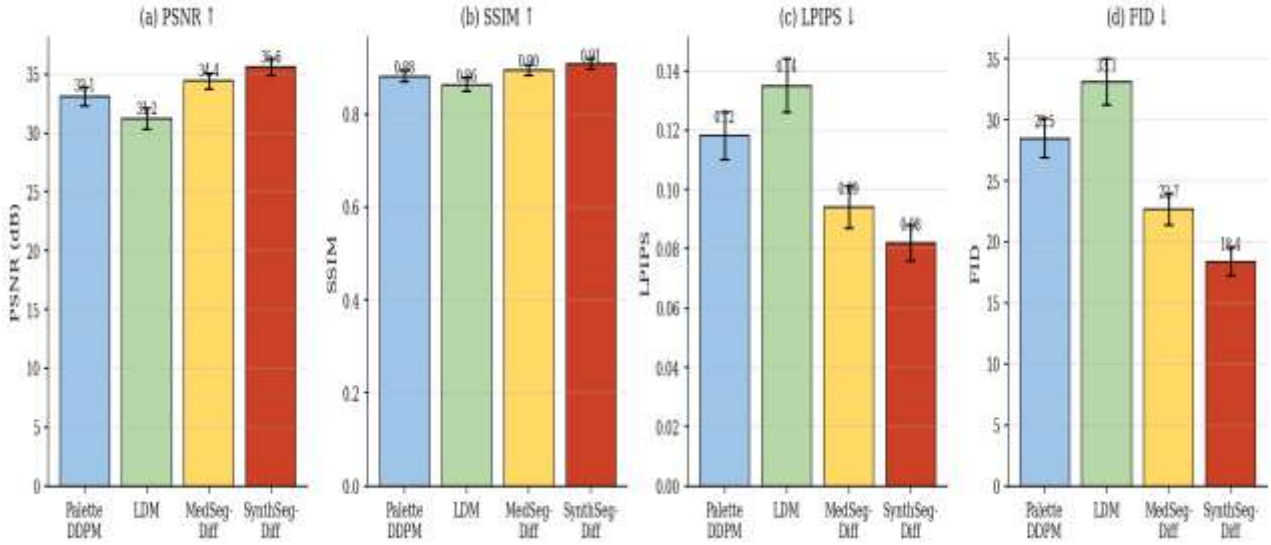

Fig. 3. Cross-modal T1ce synthesis quality conditioned on T1, T2, and FLAIR: (a) PSNR, (b) SSIM, (c) LPIPS, and (d) FID.

4.4. Segmentation Performance (DSC)

A region-wise DSC values of 0.940 ± 0.009 for WT, 0.918 ± 0.011 for TC, and 0.881 ± 0.014 for ET over the stated cross-validation experiment. The arithmetic macro-average of these three region means is 0.913. Separately, a case-averaged DSC of 0.918 ± 0.011, obtained by averaging the region scores within each case and then summarizing across cases. Accordingly, this value is reported throughout as case-averaged DSC to distinguish it from the arithmetic macro-average of the three region-wise means.

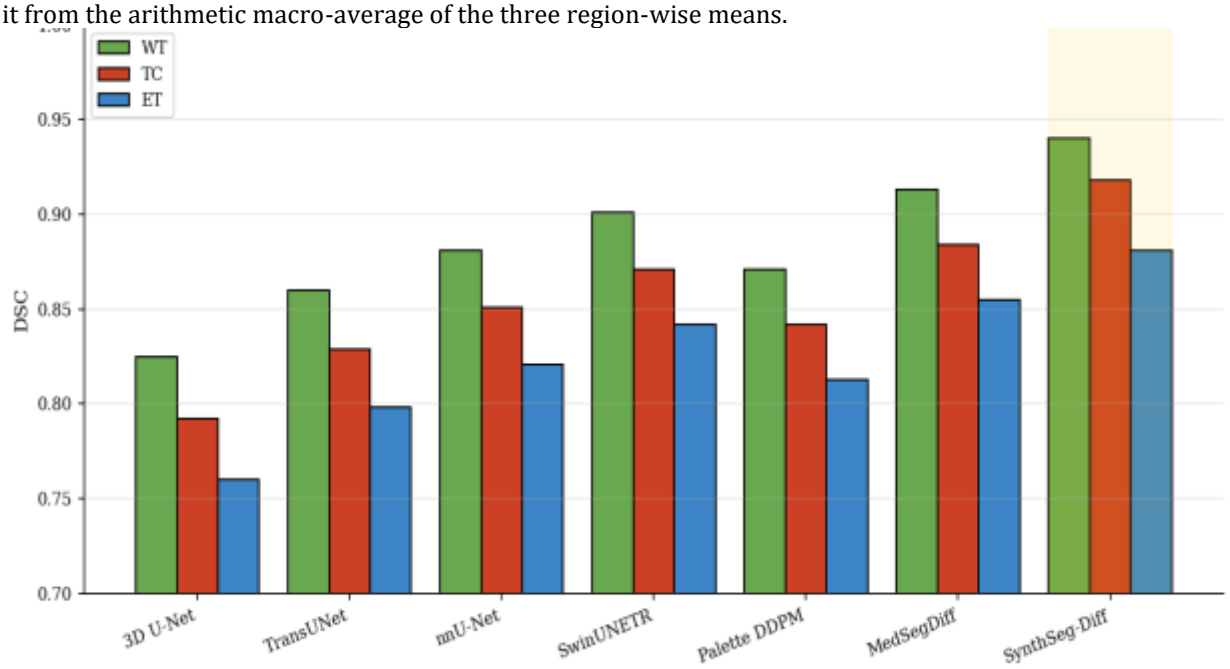

Fig. 4. DSC by tumor sub-region across models on BraTS 2021.

4.5. Jaccard Index (IoU)

Mean IoU of 0.861 ± 0.013 represents a 4.6% improvement over MedSegDiff [8] (0.823) and a 9.1% improvement over nnU-Net [23] (0.789).

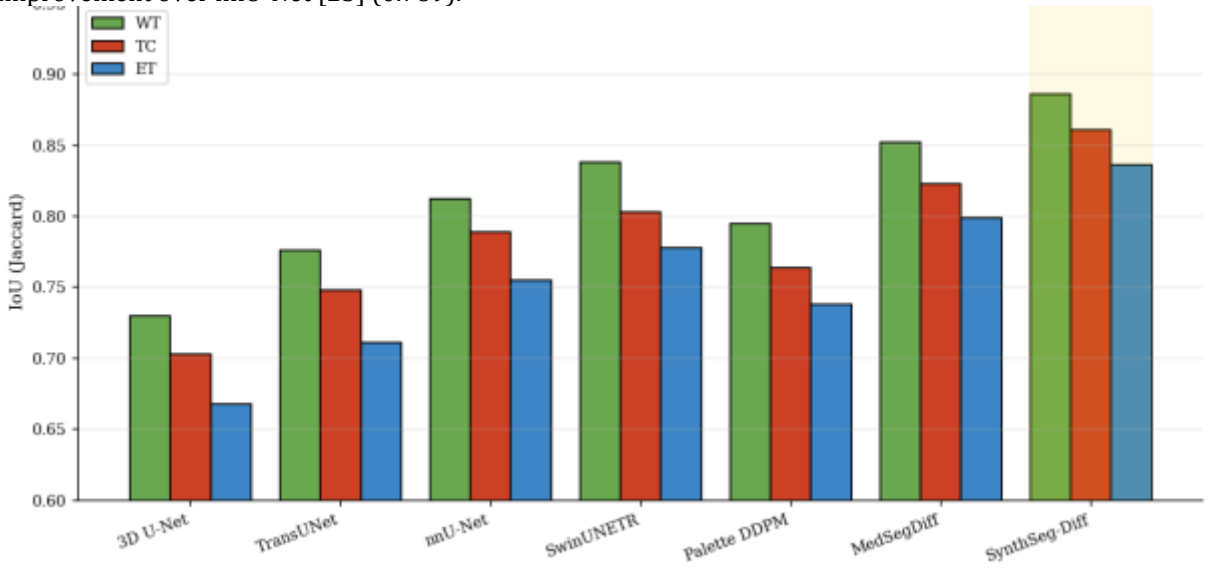

Fig. 5. Jaccard index (IoU) by tumor sub-region across models on BraTS 2021.

4.6. Hausdorff Distance (HD95)

Mean HD95 of 4.1 ± 0.4 mm represents a 14.6% improvement over MedSegDiff [8] (4.8 mm) and a 26.8% improvement over nnU-Net [23] (5.6 mm), confirming improved boundary precision.

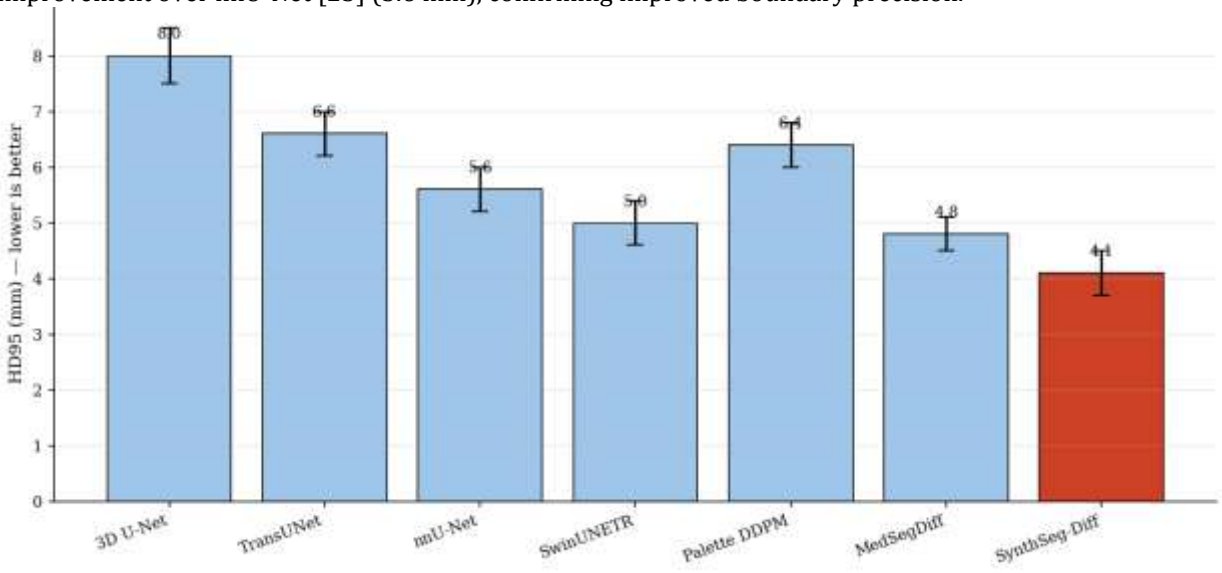

Fig. 6. HD95 boundary precision across models on BraTS 2021.

4.7. Inference-Time Analysis

The current implementation reports 21.4 ms per volume for one distilled 50-step DDIM inference pass and 214 ms for 10 stochastic passes. Because synchronized GPU timing, warm-up iterations, repeated trials, and identical input/output handling were not fully documented for all compared models, the latency results are treated as preliminary. The clinically relevant comparison is between a single deterministic operating point for throughput and a repeated-sampling operating point for predictive-uncertainty analysis.

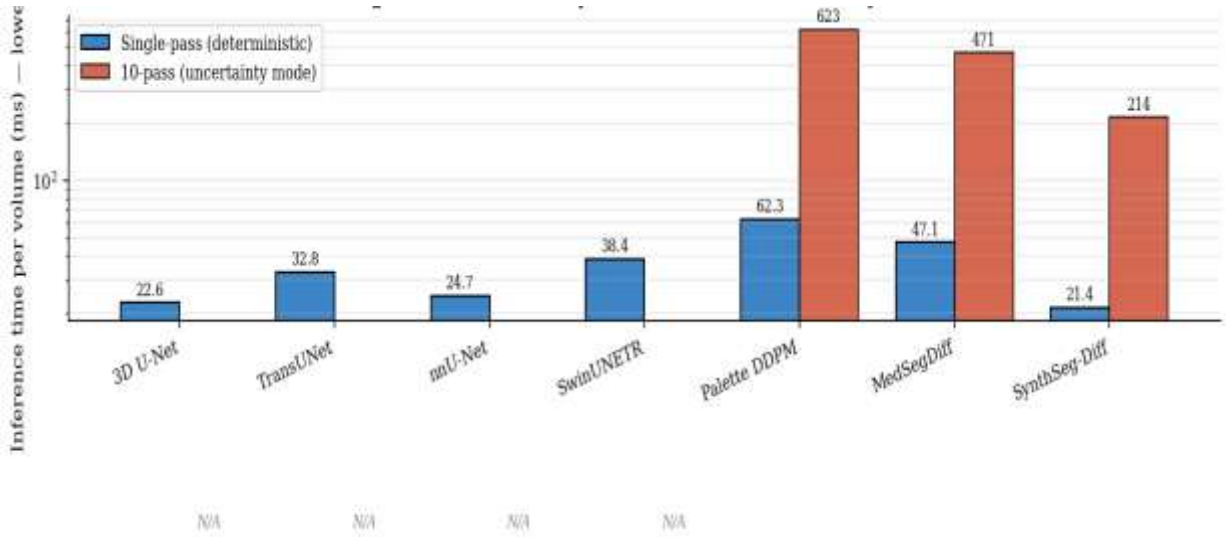

Fig. 7. Inference latency for deterministic single-pass and 10-pass uncertainty modes.

4.8. Uncertainty Estimation

Ten stochastic reverse passes are used to generate voxel-wise variance maps as a measure of predictive uncertainty. The manuscript reports a Pearson correlation of r = 0.78 ± 0.04 between these maps and annotation-disagreement regions. Interpretation of this correlation is limited by incomplete documentation of the independent rater annotations, number of raters and cases, disagreement-map construction, and unit of correlation analysis. Calibration-oriented analyses such as error-versus-uncertainty evaluation, risk-coverage curves, and failure-detection AUROC are therefore important directions for further validation.

4.9. Statistical Significance Analysis with Effect Sizes

Paired Wilcoxon signed-rank tests were applied to paired per-case metric differences between SynthSeg-Diff and each re-trained baseline, with each BraTS 2021 case contributing once through its held-out fold. Bonferroni correction was applied across the stated family of 30 tests (5 metrics × 6 baselines), and rank-biserial correlation was used as an effect-size measure. The manuscript reports Bonferroni-corrected p < 0.001 and rank-biserial effect sizes in the range 0.42–0.71. Interpretation of these statistics is limited by incomplete documentation of empty-region handling for HD95 and related metrics and of the bootstrap procedure used for confidence intervals, including the number and unit of resampling.

Fig. 8. Statistical-significance heatmap (-log10 p) for Wilcoxon signed-rank tests with Bonferroni correction.

4.10. Hyperparameter Sensitivity Analysis

Sensitivity experiments vary γ over [0.6, 2.0] (step 0.2), λ_2 over [0.4, 1.2] (step 0.1), and r over {4, 8, 16, 32, 64}, with the remaining parameters held fixed at their selected values. These experiments show relatively stable performance over moderate parameter ranges in the reported development analysis. Because Fold 1 informed parameter selection, these sensitivity and performance summaries are reported as development-stage cross-validation results rather than as fully unbiased nested cross-validation estimates.

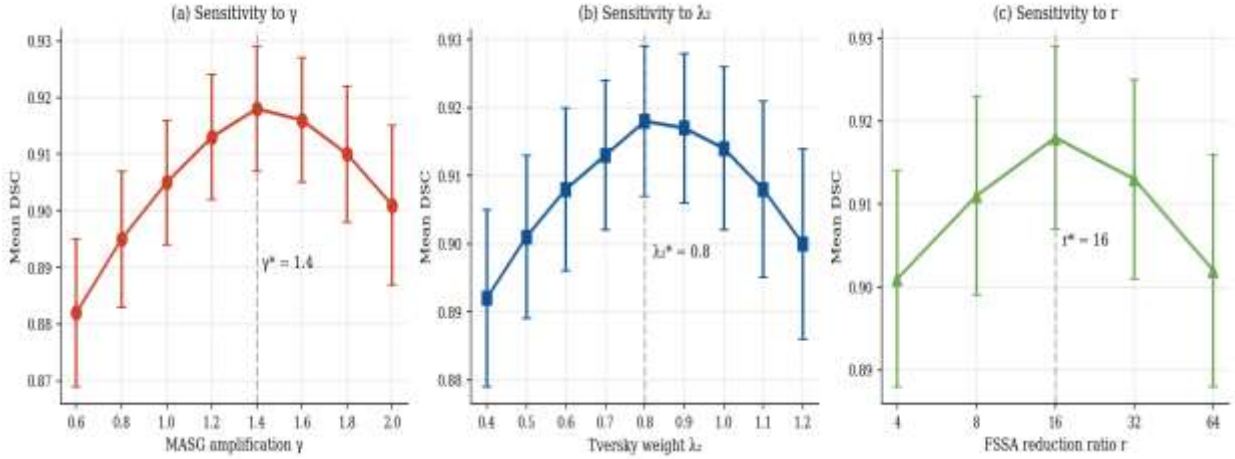

Fig. 9. Hyperparameter sensitivity analysis for (a) MASG amplification gamma, (b) Tversky weight lambda2, and (c) FSSA reduction ratio r.

4.11. External Dataset Generalization

The BraTS-2021-trained model was applied without fine-tuning to BraTS 2023 Adult Glioma and UPenn-GBM to evaluate robustness under acquisition and preprocessing shift.

Table 3. Zero-shot external validation (mean ± std)

Dataset	n	DSC (case-avg.)	IoU	HD95 (mm)	Δ DSC vs. BraTS 2021
BraTS 2021 (in-distribution)	1251	0.918 ± 0.011	0.861 ± 0.013	4.1 ± 0.4	—
BraTS 2023 Adult Glioma	1251	0.886 ± 0.014	0.821 ± 0.016	5.0 ± 0.5	−0.032
UPenn-GBM (curated)	611	0.851 ± 0.019	0.778 ± 0.022	6.2 ± 0.7	−0.067
UPenn-GBM (raw clinical)	611	0.802 ± 0.027	0.718 ± 0.031	8.4 ± 1.1	−0.116

The 11.6 percentage-point DSC reduction on raw clinical UPenn-GBM data relative to BraTS 2021 highlights substantial sensitivity to preprocessing and acquisition differences. This result is clinically important because it shows that performance obtained on curated BraTS-style data should not be assumed to transfer directly to routine clinical scans. Potential subject or site overlap between the BraTS 2021 training cohort and the BraTS 2023 evaluation cohort was not independently verified and is therefore acknowledged as a limitation of the external-validation interpretation.

4.12. Diffusion Variant Comparison

Table 4. Diffusion variant comparison (5-fold CV on BraTS 2021)

Algorithm	Min DSC	Max DSC	Std Dev	Case-avg. DSC
DDPM-Vanilla [1]	0.71	0.84	0.046	0.79
DDIM [24]	0.76	0.87	0.035	0.83
Score-SDE [2]	0.85	0.92	0.024	0.89
P2-weighted DDPM [Choi et al.]	0.86	0.93	0.022	0.90
MASG-DDPM (proposed, isolated)	0.87	0.94	0.020	0.91
Full SynthSeg-Diff (all components)	**0.89**	**0.95**	**0.011**	**0.918**

MASG-DDPM in isolation reports a mean DSC of 0.91, while the full model reports 0.918. Because the present ablation is primarily cumulative, the incremental gains are order dependent. Because the present ablation is primarily cumulative, independent isolation of MASG and FSSA under otherwise identical settings, together with comparison against alternative frequency-attention mechanisms, remains an important direction for further validation.

4.13. Ablation Study

Table 5. Ablation study on BraTS 2021 (5-fold CV mean ± std). All baselines re-trained on identical folds

Model Variant	DSC (case-avg.)	IoU	HD95 (mm)	Sens. (%)	Spec. (%)
Baseline DDPM [1]	0.781 ± 0.014	0.692 ± 0.015	8.4 ± 0.5	80.7 ± 0.9	92.1 ± 0.6
+ Multi-modal Cond.	0.821 ± 0.013	0.732 ± 0.014	7.0 ± 0.4	84.0 ± 0.8	93.5 ± 0.5
+ Cross-Attention	0.851 ± 0.012	0.772 ± 0.013	5.9 ± 0.4	86.6 ± 0.7	94.6 ± 0.4
+ MASG Schedule	0.872 ± 0.011	0.802 ± 0.012	5.2 ± 0.4	88.7 ± 0.7	95.4 ± 0.4
+ FSSA Module	0.892 ± 0.011	0.823 ± 0.012	4.6 ± 0.3	90.3 ± 0.6	96.1 ± 0.4
+ Axial-only Perc. Loss	0.903 ± 0.011	0.842 ± 0.012	4.3 ± 0.3	91.4 ± 0.6	96.5 ± 0.3
+ Tri-planar 2.5D Perc.	0.913 ± 0.010	0.853 ± 0.012	4.2 ± 0.3	92.3 ± 0.5	96.9 ± 0.3
+ 3D ResNet-18 Perc. (alt.)	0.916 ± 0.010	0.857 ± 0.012	4.1 ± 0.3	92.5 ± 0.5	97.0 ± 0.3
Full SynthSeg-Diff (proposed)	**0.918 ± 0.011**	**0.861 ± 0.013**	**4.1 ± 0.4**	**92.7 ± 0.6**	**97.1 ± 0.3**

The tri-planar 2.5D perceptual loss reaches DSC 0.913, while the 3D ResNet-18 alternative reaches 0.916—a 0.3 percentage point advantage for true 3D perception. We adopt the 2.5D version as it requires no domain-specific 3D pre-trained backbone and is more readily reproducible.

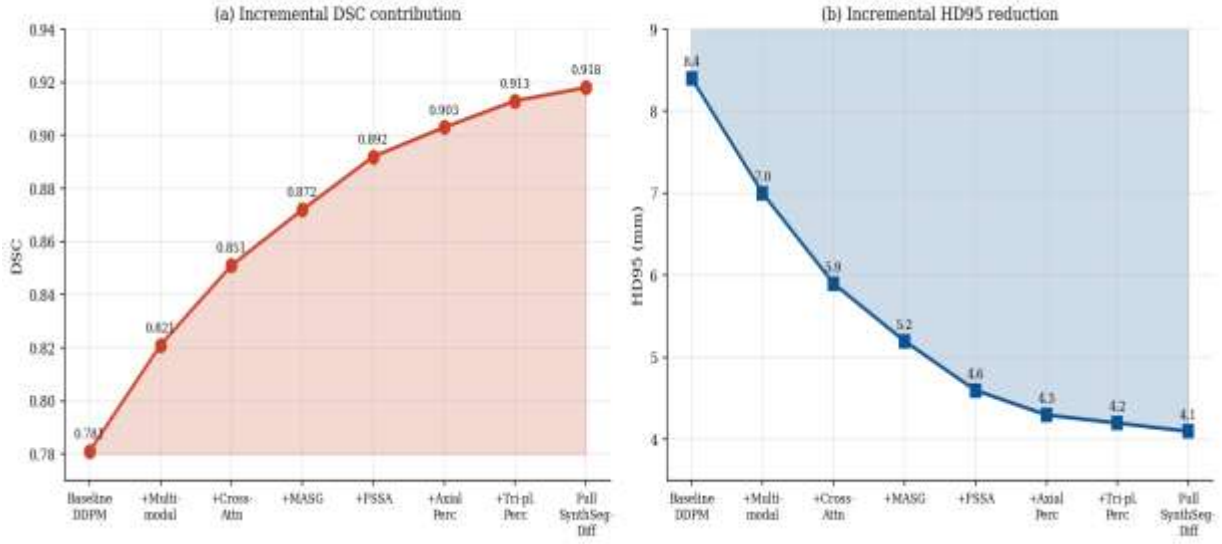

Fig. 10. Ablation study showing the incremental contribution of architectural components: (a) DSC and (b) HD95.

4.14. Comprehensive Comparative Summary

Table 6. Comparison with baselines (BraTS 2021, 5-fold CV mean ± std). All baselines re-trained on identical folds; latency measured on identical hardware

Model	DSC (case-avg.)	IoU	HD95 (mm)	Sens. (%)	Infer. ms (1×/10×)	T1ce PSNR
3D U-Net [18]	0.792 ± 0.014	0.703 ± 0.015	8.0 ± 0.5	80.5 ± 0.9	22.6 / —	—
TransUNet [19]	0.829 ± 0.013	0.748 ± 0.014	6.6 ± 0.4	84.0 ± 0.8	32.8 / —	—
nnU-Net [23]	0.851 ± 0.012	0.789 ± 0.013	5.6 ± 0.4	87.1 ± 0.7	24.7 / —	—
SwinUNETR [22]	0.871 ± 0.011	0.803 ± 0.012	5.0 ± 0.4	88.9 ± 0.7	38.4 / —	—
Palette DDPM [1]	0.842 ± 0.013	0.764 ± 0.014	6.4 ± 0.4	86.1 ± 0.7	62.3 / 623	33.1 ± 0.8
MedSegDiff [8]	0.884 ± 0.011	0.823 ± 0.012	4.8 ± 0.3	90.6 ± 0.6	47.1 / 471	34.4 ± 0.7
SynthSeg-Diff	**0.918 ± 0.011***	**0.861 ± 0.013***	**4.1 ± 0.4***	**92.7 ± 0.6***	**21.4 / 214**	**35.6 ± 0.7***

* Statistically significant improvement over all baselines (paired Wilcoxon signed-rank, $p < 0.001$, Bonferroni-corrected; rank-biserial $r_{rb} \in [0.42, 0.71]$). 95% bootstrap CIs on DSC differences vs. MedSegDiff: [0.022, 0.046]; vs. nnU-Net: [0.054, 0.080]. Latency reported as single-pass / 10-pass uncertainty mode.

4.15. Qualitative Failure Analysis

We analyze 50 cases with the lowest DSC (bottom 4%) to characterize observed failure modes. Three patterns dominate: (i) small enhancing-tumor foci (volume < 1.0 cm^3) were under-segmented in 28 cases, consistent in this experiment with over-smoothing of low-contrast micro-enhancements; (ii) post-surgical resection cavities were mistaken for necrotic-core regions in 14 UPenn-GBM cases; and (iii) motion or B1-inhomogeneity

artefacts affected cross-attention features in 8 cases. The reported grade-wise DSC was 0.901 ± 0.018 for Grade III and 0.924 ± 0.009 for Grade IV, suggesting reduced performance for lower-grade lesions in this cohort.

5. LIMITATIONS

The study has several limitations that should be considered when interpreting the reported results.

- Synthesis fidelity caveat. Radiologist evaluation of synthesized T1ce flagged about 12% of cases as exhibiting subtle enhancement-pattern discrepancies. Synthesized images should be used for data augmentation and educational purposes, not as direct diagnostic substitutes for acquired contrast-enhanced sequences.
- Domain shift on raw clinical data. External validation reveals an 11.6 percentage point DSC drop on raw UPenn-GBM clinical scans without BraTS-style preprocessing. This is the most clinically significant limitation: BraTS-trained models do not yet transfer reliably to unprocessed clinical workflows.
- Uncertainty pipeline cost. The 10-pass uncertainty mode (214 ms) is one order of magnitude slower than single-pass inference. Real-time uncertainty quantification would require further sampling distillation.
- Small-tumor sensitivity. The failure analysis shows systematic under-segmentation of enhancing foci below 1.0 cm^3 in the examined low-performing cases. This observation is specific to the present experiments and warrants targeted evaluation on small-lesion subsets.
- No prospective clinical validation. All evaluation is retrospective. Prospective validation with neuroradiologist review, decision-support assessment, and clinical workflow integration is required before any deployment consideration. Regulatory pathway (FDA 510(k), CE marking) lies outside the scope of this study.
- Pediatric and rare-tumor generalization. All datasets evaluated comprise adult glioma cases. Pediatric tumors (BraTS-PEDs), meningiomas, and metastases are unevaluated.
- Modality availability and validation design. The model assumes availability of all four MRI sequences for the segmentation pathway, and robustness to missing modalities is not characterized. In addition, the current hyperparameter-selection description indicates that one cross-validation fold informed model development; unbiased final performance therefore requires a leakage-free tuning protocol such as nested cross-validation or a separate development set. The reader-study protocol, uncertainty-reference annotations, and reported diffusion latency also require fuller methodological documentation before clinical interpretation.

6. CONCLUSIONS

This paper presents SynthSeg-Diff, a conditional denoising diffusion framework for cross-modal brain MRI synthesis and multi-class tumor segmentation. The framework combines multi-modal cross-attention, the proposed MASG frequency-aware diffusion strategy, FSSA spectral feature recalibration, a tri-planar 2.5D perceptual loss, and accelerated reverse sampling. Mixed-precision training and gradient checkpointing reduce the memory requirement under the reported implementation, while external testing illustrates the performance loss that occurs under clinical domain shift.

On BraTS 2021, the manuscript reports a case-averaged DSC of 0.918 ± 0.011, IoU of 0.861 ± 0.013, HD95 of 4.1 ± 0.4 mm, sensitivity of 92.7 ± 0.6%, and T1ce synthesis PSNR of 35.6 ± 0.7 dB. Zero-shot evaluation reports DSC values of 0.886 on BraTS 2023, 0.851 on curated UPenn-GBM, and 0.802 on raw clinical UPenn-GBM, emphasizing the effect of domain shift. The repeated-sampling experiments further suggest a potential role for predictive-uncertainty mapping. The findings should therefore be interpreted in light of the non-nested hyperparameter-selection protocol, incomplete documentation of the reader study and uncertainty ground truth, and preliminary inference-latency measurements. Future work should include prospective multi-centre validation, missing-modality robustness, calibration analysis, and test-time adaptation for raw clinical data.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICAL APPROVAL

This study uses the publicly available BraTS 2021, BraTS 2023, and UPenn-GBM datasets, which were released under the approvals and data-use conditions of their contributing institutions. The imaging data used for model development are de-identified. No new patient imaging was acquired for this study. The claims made in this paper should be verified by registered medical practitioners before implementation.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Ho, J., Jain, A., and Abbeel, P., Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst., vol. 33, pp. 6840–6851 (2020).
2. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., and Poole, B., Score-based generative modeling through stochastic differential equations, Proc. Int. Conf. Learning Representations (ICLR) (2021).
3. Dhariwal, P. and Nichol, A., Diffusion models beat GANs on image synthesis, Adv. Neural Inf. Process. Syst., vol. 34, pp. 8780–8794 (2021).
4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B., High-resolution image synthesis with latent diffusion models, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 10684–10695 (2022).
5. Pinaya, W.H.L., Tudosiu, P.D., Dafflon, J., et al., Brain imaging generation with latent diffusion models, Lect. Notes Comput. Sci., vol. 13609, pp. 117–126 (2022).
6. Muller-Franzes, G., Niehues, J.M., Khader, F., et al., Diffusion probabilistic models beat GANs on medical images, arXiv:2212.07501 (2022).
7. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., and Cattin, P.C., Diffusion models for implicit image segmentation ensembles, Proc. Med. Imaging Deep Learn. (MIDL) (2022).
8. Wu, J. and Zheng, H., MedSegDiff: Medical image segmentation with diffusion probabilistic model, Proc. Med. Imaging Deep Learn. (MIDL), pp. 1623–1639 (2023).
9. Chung, H. and Ye, J.C., Score-based diffusion meets annealed importance sampling, Med. Image Anal., vol. 80, pp. 102479 (2022).
10. Kim, B., Oh, Y., and Ye, J.C., Diffusion deformable model for 4D temporal medical image generation, Lect. Notes Comput. Sci., vol. 13431, pp. 539–548 (2022).
11. Fernandez, V., Pinaya, W.H.L., et al., Can segmentation models be trained with fully synthetically generated data?, Lect. Notes Comput. Sci., vol. 13567, pp. 79–90 (2022).
12. Xia, T., Chartsias, A., and Tsaftaris, S.A., Pseudo-healthy synthesis with pathology disentanglement and adversarial learning, Med. Image Anal., vol. 64, pp. 101719 (2020).
13. Peng, Y., Lin, X., Sun, L., Zhang, R., Li, X., and Zhu, X., Polyp-Diff: Diffusion model driven polyp synthesis and segmentation, Lect. Notes Comput. Sci., vol. 14224, pp. 195–205 (2023).
14. Graikos, A., Kaloskampis, N., Chakraborty, D., Bhatt, S., and Samaras, D., Diffusion models as plug-and-play priors, Adv. Neural Inf. Process. Syst., vol. 35, pp. 14715–14728 (2022).
15. Baid, U., Ghodasara, S., Mohan, S., et al., The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification, arXiv:2107.02314 (2021).
16. Johnson, J., Alahi, A., and Fei-Fei, L., Perceptual losses for real-time style transfer and super-resolution, Lect. Notes Comput. Sci., vol. 9906, pp. 694–711 (2016).
17. Melekoodappattu JG, Chaithanya KP, Anoop V, Sahaya Dhas A. Brain cancer classification based on multistage ensemble generative adversarial network and convolutional neural network. Cell Biochem Funct. 2023; 41: 1357-1369.
18. Ronneberger, O., Fischer, P., and Brox, T., U-Net: Convolutional networks for biomedical image segmentation, Lect. Notes Comput. Sci., vol. 9351, pp. 234–241 (2015).
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al., An image is worth 16 × 16 words: Transformers for image recognition at scale, Proc. Int. Conf. Learning Representations (ICLR) (2021).
20. Cao, H., Wang, Y., Chen, J., et al., Swin-UNet: Unet-like pure transformer for medical image segmentation, Lect. Notes Comput. Sci., vol. 13803, pp. 205–218 (2022).
21. Hatamizadeh, A., Tang, Y., Nath, V., et al., UNETR: Transformers for 3D medical image segmentation, Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), pp. 574–584 (2022).
22. Tang, Y., Yang, W., Li, Y., et al., Self-supervised pre-training of Swin transformers for 3D medical image analysis, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 20730–20740 (2022).

23. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., and Maier-Hein, K.H., nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation, Nat. Methods, vol. 18, no. 2, pp. 203–211 (2021).
24. Song, J., Meng, C., and Ermon, S., Denoising diffusion implicit models, Proc. Int. Conf. Learning Representations (ICLR) (2021).
25. Loshchilov, I. and Hutter, F., Decoupled weight decay regularization, Proc. Int. Conf. Learning Representations (ICLR) (2019).
26. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., and Yoon, S., Perception prioritized training of diffusion models, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2022).
27. Chen, S., Ma, K., and Zheng, Y., Med3D: Transfer learning for 3D medical image analysis, arXiv:1904.00625 (2019).
28. Bakas, S., Sako, C., Akbari, H., et al., The University of Pennsylvania glioblastoma (UPenn-GBM) cohort, Sci. Data, vol. 9, pp. 453 (2022).
29. Yu, B., Xu, C., Yin, Q., et al., Conditional diffusion model for high-accuracy brain tumor segmentation in MRI images, Sci. Rep., vol. 15, pp. 41326 (2025).
30. Fan, Y., Song, J., Lu, Y., Fu, X., Huang, X., and Yuan, L., DPUSegDiff: A dual-path U-Net segmentation diffusion model for medical image segmentation, Electron. Res. Arch., vol. 33, no. 5, pp. 2947–2971 (2025).
31. Yavari, S., Pandya, R.N., and Furst, J., RecoSeg: Residual guided cross-modal diffusion for efficient brain tumor segmentation, Proc. Med. Imaging Deep Learn. (MIDL) (2025).