Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

Anatomically-conditioned Latent Diffusion Model for Data-Efficient Few-Shot Cross-Domain 3D Glioma MRI Synthesis

Salman Shaik[†,*], Truong Thanh Hung Nguyen[†,*], Hung Cao[†]
[†]Analytics Everywhere Lab, University of New Brunswick, Canada

Abstract

Accurate classification of diffuse gliomas is often hindered by domain shifts across centers and a lack of large, annotated datasets. We propose the *Anatomically-conditioned Latent Diffusion Model (ALDM)*, a novel framework for data-efficient, few-shot 3D volumetric MRI synthesis. ALDM utilizes a two-stage approach: a 3D variational autoencoder learns anatomical priors from a data-rich source domain, while a conditional latent diffusion model, guided by tumor masks via a ControlNet, generates structurally coherent volumes for a data-scarce target domain. Evaluated in an extreme few-shot setting with only 16 target images, ALDM outperformed GAN and hybrid baselines, achieving a superior Fréchet Inception Distance (FID) of 85.40 and a downstream classification AUC of 0.987. Qualitative results confirm that the model preserves sharp pathology boundaries and cross-modal consistency, with visual fidelity improving progressively during training. By capturing essential diagnostic features, ALDM provides a robust tool for clinical data augmentation in low-resource settings. Our implementation is available at https://github.com/Analytics-Everywhere-Lab/anatomically-conditioned-LDM.

Keywords: Few-shot Cross-domain Image Synthesis, Latent Diffusion Models, Generative AI, 3D Glioma MRI

1. Introduction

Diffuse gliomas are heterogeneous malignant brain tumors with varied prognoses, so accurate classification of grade and molecular subtype (e.g., IDH mutation, 1p19q codeletion) is important for treatment planning [1, 2]. Deep learning (DL) models can classify gliomas from multimodal MRI with high accuracy on single-institution data, but performance often drops under multicenter domain shift and in limited-sample settings due to the lack of large, consistently annotated datasets [3–6]. Radiological heterogeneity further complicates generalization. WHO Grade IV glioblastoma, particularly IDH-wildtype, often shows strong contrast enhancement and elevated cerebral blood volume, whereas WHO Grade II-III and IDH-mutant gliomas more often exhibit weaker enhancement and reduced perfusion [7]. These IDH-linked differences induce a clinically meaningful domain shift that single-domain synthesis methods do not adequately address.

Acquiring high-quality multimodal brain MRI is resource-intensive because protocols require multiple sequences and one hour of scan time. In routine practice, some sequences are missing due to limited time or image artifacts. Across institutions, differences in scanner vendor, field strength, coils, and protocol settings introduce site effects that shift image statistics and can reduce the external performance of DL models [6]. Generative Adversarial Network (GAN) and conditional Latent Diffusion Models (LDM) have shown strong synthesis quality and better diversity than GANs, and data-efficient fine-tuning can adapt diffusion generators using limited target data [8, 9]. However, most glioma synthesis studies focus on two-dimensional (2D) slices and are rarely evaluated in few-shot cross-domain settings [10, 11]. High-quality target-domain volumetric synthesis could reduce the need for additional scans and help mitigate dataset shift across sites.

[*]salmanbasha.shaik@unb.ca, hung.ntt@unb.ca

This motivates the proposed *Anatomically-conditioned Latent Diffusion Model (ALDM)*, a novel framework for three-dimensional (3D) volumetric glioma MRI synthesis across heterogeneous source and target glioma populations in few-shot learning scenarios. ALDM addresses a significant research gap by integrating latent diffusion processes with anatomical structure conditioning to achieve data-efficient, spatially coherent synthesis optimized for downstream glioma classification tasks. By synthesizing high-quality target-domain volumes from limited source-domain data, ALDM enables practical data augmentation for glioma datasets while preserving the domain-specific imaging characteristics essential for classification accuracy. The main contributions are summarized as follows:

(1) We propose ALDM, which combines a 3D variational autoencoder (VAE) for source-domain anatomical representation learning with a conditional LDM guided by tumor masks, enabling controlled synthesis of multimodal volumetric glioma MRI across heterogeneous source and target domain imaging protocols, generating anatomically coherent T1-weighted (T1), T2-weighted (T2), and T2-weighted fluid-attenuated inversion recovery (FLAIR) images with preserved tumor heterogeneity and domain-specific imaging characteristics.

(2) We show ALDM synthesizes anatomically consistent 3D MRI volumes when transferring from a data-rich and homogeneous glioblastoma (GBM) source domain [2] to a data-scarce and heterogeneous pre-operative diffuse glioma (PDGM) target domain [1] under few-shot and consistently outperform GAN-based and hybrid baselines in both image-level similarity metrics and downstream classification performance.

2. Related Work

2.1. Domain Heterogeneity in Glioma MRI Datasets

Domain heterogeneity in glioma MRI arises from both biological variability and acquisition differences across sites. Public datasets such as BraTS and TCGA/TCIA aggregate scans acquired with different scanner vendors, field strengths, coils, reconstruction pipelines, and protocol settings, which introduce systematic shifts in resolution, noise, and contrast even after curation and preprocessing [12, 13].

These site effects create a distribution mismatch between development and deployment data and can substantially degrade out-of-domain performance [14]. In glioma segmentation, models that perform well on curated challenge data often drop in accuracy on routine clinical MRI, where image quality is more variable, and modality sets may be incomplete. External multi-center studies report clear gaps between BraTS-style evaluation and real hospital data, and suggest improved robustness when training includes more heterogeneous cohorts or explicitly handles missing modalities [15]. More generally, deep models may rely on scanner- or protocol-specific cues correlated with labels in the training set, reducing robustness under shifts in hardware, pulse sequences, or preprocessing [6].

To improve cross-site robustness, common approaches include statistical harmonization that models sites as batch effects [5], domain adaptation that encourages domain-invariant representations [4], and generative harmonization or synthesis for scanner-style mapping and missing-sequence completion [3]. Despite progress, reliable generalization without target labels or site-specific retraining remains difficult, especially when protocols and patient populations differ substantially across institutions.

2.2. Generative Models for Volumetric Glioma MRI Synthesis

Data scarcity and class imbalance in brain tumor cohorts have driven the development of generative models for MRI augmentation. While early GAN-based methods primarily synthesized 2D slices, recent approaches such as AGGrGAN [11] and LASTGAN [10] have

improved fidelity and domain specificity by incorporating multi-component aggregation and style encoders. Generating full 3D tumor volumes is more challenging because models must preserve spatial consistency across depth. Recent work has therefore moved toward volumetric synthesis. A 3D VQ-GAN combined with a Transformer has been proposed to generate high-resolution tumor ROIs within MRI volumes [16]. Diffusion models have also shown strong fidelity and diversity for 3D medical synthesis. Med-DDPM uses semantic conditioning within a 3D diffusion framework and achieves segmentation performance close to that on real data when synthetic volumes are used for augmentation [8]. Similarly, mask-conditioned latent diffusion with a 3D autoencoder enables multi-contrast tumor volume generation from tumor masks while maintaining anatomical consistency [9].

Despite these advances, data-efficient 3D glioma MRI synthesis remains insufficiently studied. Most 3D GAN or diffusion pipelines still rely on moderate training set sizes, and few-shot settings with strong cross-domain shift have not been systematically validated. This gap motivates the development of generative frameworks that can learn from very limited target data while producing anatomically plausible and structurally consistent 3D MRI tumor volumes.

3. Architecture

The proposed ALDM is a conditional latent-diffusion-based framework designed for cross-domain synthesis of 3D MRI volumes under data scarcity. As illustrated in Figure 1, the architecture decomposes the generative process into two stages: (i) source-domain anatomical representation learning using a VAE, and (ii) cross-domain conditional latent diffusion generation in the learned latent space.

Problem Formulation

Let the source (data-rich, homogeneous) domain be $\mathcal{D}_s = \{(x_i^s, m_i^s)\}_{i=1}^{N_s}$ and the target (data-scarce, heterogeneous) domain be $\mathcal{D}_t = \{(x_j^t, m_j^t)\}_{j=1}^{N_t}$, where $x \in \mathbb{R}^{C\times D\times H\times W}$ denotes a multi-modal 3D MRI volume and m is the corresponding tumor segmentation mask. Here, $C = 3$ corresponds to T1, T2, and FLAIR, and (D, H, W) are the volumetric spatial dimensions. Our goal is to learn a conditional generative model for the target domain that synthesizes a realistic volume $\hat{x}^t$ given an anatomical prior derived from the mask. We denote this prior as $c = f(m)$, where $f(\cdot)$ may include the binary mask and auxiliary mask-derived cues (e.g., edges or distance transforms). Generation is performed from a latent variable z such that: $\hat{x}^t = G(z, c)$.

We consider the *few-shot* setting where $N_t = K \ll N_s$, with K labeled target-domain examples available for adaptation. The challenge is to preserve shared organ-level anatomy across domains while allowing controlled variation in tumor appearance guided by c.

3.1. Source-Domain Anatomical Representation Learning

In the first stage, a 3D VAE is trained on the data-rich GBM MRI dataset to learn a compact and anatomically consistent latent representation of volumetric brain images. Given an input MRI volume $\mathbf{x} \in \mathbb{R}^{C\times D\times H\times W}$, where $C = 3$ corresponds to the T1, T2, and FLAIR modalities and $(D, H, W) = (112, 112, 112)$ after preprocessing, the VAE encoder $q_\phi(\mathbf{z} \mid \mathbf{x})$ compresses the high-resolution input into a lower-dimensional latent tensor $\mathbf{z} \in \mathbb{R}^{z_C\times d\times h\times w}$ with $z_C = 8$ and $(d, h, w) = (28, 28, 28)$. This compression significantly reduces spatial complexity while preserving global anatomical structure, making subsequent generative modeling more stable and memory efficient.

The encoder is implemented as a stack of 3D convolutional blocks with kernel size $3\times3\times3$. Two strided convolution layers perform spatial downsampling by a factor of 4, yielding

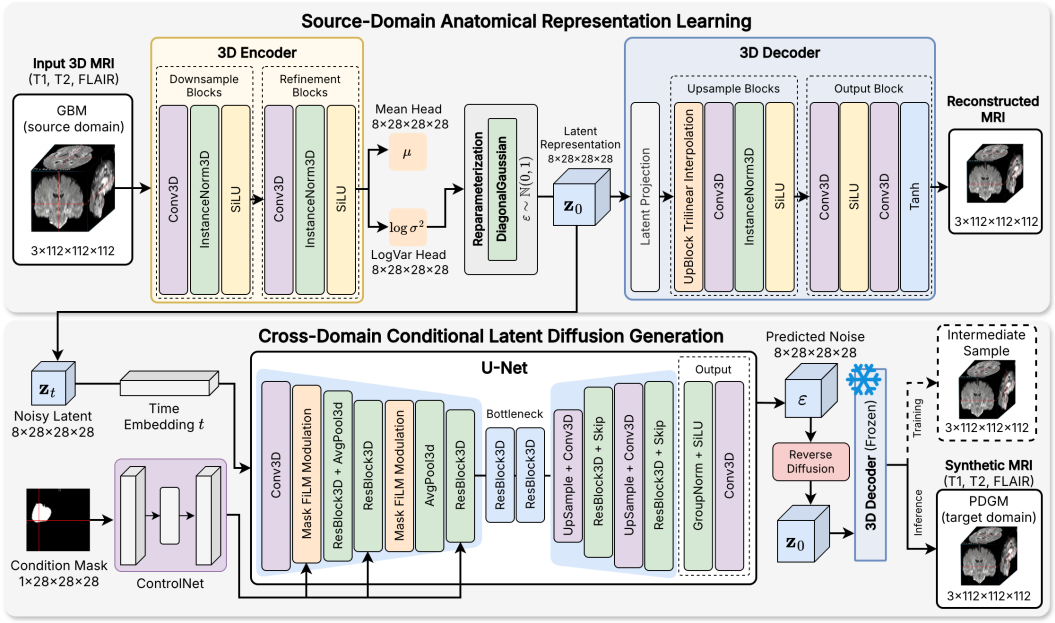

Figure 1. Overview of the proposed Anatomically-conditioned Latent Diffusion Model (ALDM) for cross-domain MRI synthesis. A VAE encoder compresses 3D GBM MRI volumes into a latent space, where a U-Net-based denoising diffusion model iteratively denoises under anatomical-mask conditioning. The decoded output produces synthetic PDGM MRI volumes.

successive spatial resolutions of 112, 56, and 28. At each stage, instance normalization and nonlinear activations are applied to ensure stable training under the small batch sizes typical of volumetric medical imaging. Additional convolutional refinement blocks are inserted at the latent resolution to increase representational capacity without further reducing spatial detail, allowing the encoder to capture fine-grained anatomical variations. The encoder predicts the parameters of a diagonal Gaussian posterior distribution as:

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\left(\boldsymbol{\mu}_\phi(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\mathbf{x}))\right), \tag{3.1}$$

from which latent samples are drawn using the reparameterization. To prevent numerical instability, the predicted log-variance is clamped to a fixed range during training.

The decoder mirrors the encoder architecture and reconstructs volumetric MRI data from latent samples. It consists of two successive upsampling stages implemented using trilinear interpolation followed by 3D convolution, resulting in spatial resolutions of 28, 56, and 112. This design avoids checkerboard artifacts commonly associated with transposed convolutions and encourages smooth volumetric reconstructions. A final tanh activation constrains voxel intensities to the normalized range $[-1, 1]$. By reconstructing full 3D volumes rather than independent slices, the VAE enforces volumetric coherence across slices and preserves consistent anatomical structure throughout the brain. The VAE is trained using a weighted combination of reconstruction and Kullback–Leibler (KL) divergence losses as:

$$\mathcal{L}_{\mathrm{VAE}} = \lambda_{\mathrm{rec}} \|\mathbf{x} - \hat{\mathbf{x}}\|_1 + \lambda_{\mathrm{KL}}, \quad \mathrm{KL}\big(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\big). \tag{3.2}$$

where the reconstruction term encourages voxel-level fidelity and the KL term regularizes the latent distribution toward a unit Gaussian prior. A deliberately small λ_{KL} is used to mitigate posterior collapse and retain anatomically meaningful variability in the latent space. To prevent posterior collapse during early training, the KL warm-up is applied. The

KL weight is linearly increased over a fraction $\rho = 0.35$ of the total training steps:

$$\lambda_{\mathrm{KL}}(t) = \begin{cases} \lambda_{\mathrm{KL}} \cdot \frac{t}{\rho T}, & t \leq \rho T, \\ \lambda_{\mathrm{KL}}, & t > \rho T, \end{cases} \tag{3.3}$$

where t denotes the current training step and T the total number of steps.

A 3D gradient consistency loss further enforces structural fidelity and encourages sharp anatomical boundaries:

$$\mathcal{L}_{\nabla} = \sum_{a \in \{x,y,z\}} \left\| \nabla_a \hat{\mathbf{x}} - \nabla_a \mathbf{x} \right\|_1 . \tag{3.4}$$

After training, the encoder and decoder weights are frozen, and the encoder is used exclusively to extract latent representations for diffusion-based generation.

3.2. **Cross-Domain Conditional Latent Diffusion Generation**

In the second stage, a conditional U-Net-based Denoising Diffusion Probabilistic Model (DDPM) is trained in the VAE's latent space. Instead of operating on full-resolution voxel grids, diffusion is performed on normalized latent tensors, which significantly reduces computational cost while preserving global anatomical structure. This design is particularly important for 3D MRI data, where direct diffusion in voxel space is prohibitively expensive. The latent diffusion model is trained to transform Gaussian noise into anatomically plausible latent representations, which are subsequently decoded into synthetic MRI volumes.

3.2.1. **Forward Diffusion in Latent Space**

Forward diffusion process gradually corrupts a clean latent representation $\mathbf{z}_0$ by injecting Gaussian noise over T timesteps. At each timestep t, the noisy latent $\mathbf{z}_t$ is obtained as:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\, \mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t}\, \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3.5}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$, and $\{\beta_t\}_{t=1}^{T}$ follows a linear schedule from 10^{-4} to 2×10^{-2} over $T = 1000$ timesteps. This process progressively removes structural information from the latent representation, yielding a sequence of increasingly noisy latents that serve as training inputs for the reverse denoising model. Operating in the VAE latent space ensures that the forward diffusion primarily perturbs semantically meaningful anatomical features rather than low-level voxel noise.

3.2.2. **Reverse Denoising with a Conditional 3D U-Net**

The reverse diffusion process is parameterized by a 3D U-Net, denoted as $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})$, which is trained to predict the noise injected at each timestep. The U-Net follows an encoder–decoder architecture with skip connections and processes latent tensors at spatial resolutions of 28, 14, and 7, before symmetrically restoring the resolution to 28. Each resolution level consists of residual 3D convolutional blocks with group normalization, which provides stable training under the small batch sizes imposed by volumetric MRI data.

Temporal information is incorporated via sinusoidal timestep embeddings as:

$$\gamma(t) = [\cos(\omega_k t), \sin(\omega_k t)]_{k=1}^{d/2}, \tag{3.6}$$

which are projected through multilayer perceptrons and injected into each residual block as feature-wise biases. This conditioning allows the network to adapt its denoising behavior across diffusion steps and effectively model the time-dependent structure of the reverse process. By predicting the noise component, the model employs standard DDPM formulation and benefits from improved training stability.

3.2.3. **Conditional Generation via Latent Diffusion**

To enable cross-domain adaptation and explicit anatomical guidance, the latent diffusion model is conditioned on tumor segmentation masks derived from both the source and target domains. Conditioning is incorporated through complementary mechanisms operating at multiple spatial scales.

First, a lightweight FiLM-style modulation applies mask-dependent affine transformations to intermediate feature maps, providing a global bias toward tumor-aware representations without significantly increasing model complexity as follows:

$$\mathbf{h}' = \mathbf{h} \odot (1 + 0.1\,\gamma) + 0.1\,\beta, \tag{3.7}$$

where (γ, β) are predicted from global mask statistics.

Second, additional anatomical control signals, including edge maps and soft distance transforms derived from the tumor mask, are processed by a ControlNet [17]. The resulting residual feature maps are injected into the main U-Net at multiple resolutions, enforcing spatial alignment between generated latent structures and anatomical priors. This multi-scale conditioning strategy encourages the preservation of tumor location, shape, and spatial extent during generation, while still allowing the diffusion process to model realistic intensity variation and global brain anatomy. By conditioning on anatomical structure rather than domain-specific intensity statistics alone, the model effectively transfers structural knowledge from the data-rich GBM domain to the data-scarce PDGM domain. The diffusion model is trained by minimizing the mean-squared error between the true injected noise and the predicted noise with additional spatial weighting applied within tumor regions to prevent lesion attenuation during denoising.

3.2.4. **Conditional Guidance and Loss Weighting**

To enable flexible control over anatomical conditioning strength, we adopt classifier-free guidance during latent diffusion training. Specifically, conditioning inputs are randomly dropped with probability $p_{\text{drop}} = 0.1$, such that the denoising network learns both conditional and unconditional noise predictions. During inference, the guided noise estimate is computed as:

$$\boldsymbol{\epsilon}_{\text{guided}} = \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \varnothing) + s\,(\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c}) - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \varnothing))\,, \tag{3.8}$$

where $\varnothing$ denotes null conditioning and s is the guidance scale. This formulation allows explicit trade-off between anatomical adherence and sample diversity at generation time.

To further preserve tumor structure during denoising, the diffusion loss is spatially weighted using the ground-truth tumor mask. In addition to the binary tumor region, a dilated neighborhood is included to encourage smooth structural transitions at lesion boundaries. The dilation is implemented using a $3 \times 3 \times 3$ structuring element in latent space. Formally, the weighted diffusion objective is expressed as

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[w(\mathbf{m})\,\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2\right], \tag{3.9}$$

where $w(\mathbf{m})$ assigns higher weights to voxels within the tumor and its local neighborhood. The overall influence of structural control signals is modulated by a global scaling factor $\lambda_{\text{ctrl}} = 1.0$, which is applied uniformly across all conditioning pathways.

After the reverse diffusion process converges, the denoised latent representation $\mathbf{z}_0$ is passed through the frozen VAE decoder to produce a synthetic 3D MRI volume as:

$$\mathbf{x}_{\text{syn}} = \text{Dec}(\mathbf{z}_0), \quad \mathbf{z}_0 = \text{DDPM}^{-1}(\mathcal{N}(0, I), \mathbf{c}). \tag{3.10}$$

This formulation tightly couples anatomical conditioning with diffusion-based generation, enabling anatomically coherent, controllable, and data-efficient cross-domain MRI synthesis under both few-shot and zero-shot target-domain settings.

4. Experiment Setup

4.1. Dataset

Our experiment uses a GBM dataset [2] and a PDGM dataset [1], both with T1, T2, and FLAIR modalities, stored primarily in NIfTI format. The GBM dataset contains approximately 828,000 image slices derived from 3D MRI volumes of multiple subjects. In addition to imaging, the GBM dataset includes structured clinical metadata files with data types including MR, molecular test results, and demographic records. While PDGM consists of approximately 12,000 images, with accompanying metadata categories including MR, measurement, demographic, follow-up, and diagnosis. This strong imbalance between a large source-domain cohort (GBM) and a limited target-domain cohort (PDGM) makes the setting well-suited for evaluating few-shot cross-domain generative transfer. Separately, the full PDGM dataset is used to train the downstream CNN classifier for evaluation, ensuring the classifier learns target-domain decision boundaries from real target data rather than synthetic outputs.

4.2. Baselines

We compare our proposed ALDM performance against our three re-implemented representative generative baselines commonly used for medical image synthesis and cross-domain adaptation. These baselines span adversarial, multi-discriminator, and hybrid latent–adversarial paradigms, providing a comprehensive comparison across modeling strategies. (1) conditional GAN (CGAN) [18], which conditions the generator on auxiliary information to guide image synthesis across domains. (2) 3M-CGAN extension by employing an ensemble of three discriminators, each operating on three corresponding MRI modalities [19]. This multi-discriminator design encourages modality-specific realism and has been shown to improve training stability and visual fidelity in multi-channel medical imaging tasks. (3) Hybrid VAE-GAN model [20], which combines a VAE for latent-space representation learning with a conditional adversarial objective to enhance sample realism.

4.3. Evaluation Metrics

In our experiment, we conducted the performance comparison across both the data-rich GBM domain and the data-scarce PDGM domain through two standards:

(1) Image-Level Fidelity: We compute the following metrics at the patient-wise fidelity level to preserve volumetric integrity and ensure clinically meaningful evaluation that reflects real-world diagnostic workflows:
   (a) *Fréchet Inception Distance (FID)* measures distributional similarity between real and synthesized MRI volumes, with lower values indicating closer alignment with the target distribution.
   (b) *The Structural Similarity Index Measure (SSIM)* quantifies local and global structural correspondence between generated and real MRI volumes, with higher values indicating greater anatomical consistency.

   Specifically, we evaluate all models on 64 target-domain subjects, each consisting of 112 aligned axial slices per modality. For each subject, SSIM and FID are computed by aggregating metrics across the 112 generated and real slices, yielding a single fidelity score per patient. Final patient-wise results are reported as the mean across the 64 subjects.

(2) Downstream classification: To assess the diagnostic utility of the generated images, a downstream AlexNet-based classifier [21] was evaluated on synthetic MRI data after being trained exclusively on real samples from the original dataset. This evaluation

framework was used to ensure the trained CNN was adhered to classify the images with high accuracy, where downstream performance serves as a proxy for clinical usefulness. Performance is measured using balanced classification accuracy (BAcc), F1 score, and the area under the ROC curve (AUC).

5. Results

5.1. Quantitative Evaluation

The quantitative results, summarized in Table 1 and visualized in Figure 2, highlight the effectiveness of the ALDM framework in the data-scarce PDGM target domain.

5.1.1. Image-Level Fidelity

Regarding image-level fidelity, the proposed ALDM ($K = 16, s = 3.0$) achieves the lowest FID of 85.40, indicating the highest distributional similarity to the real MRI volumes. While the VAE-GAN baseline achieves a higher SSIM of 0.750, it produces a higher FID (88.18) than the ALDM variants. The results suggest that while VAE-GAN excels at local pixel-wise reconstruction, the ALDM framework better captures the global distribution and stylistic nuances of the target domain. Other baselines, such as CGAN and 3M-CGAN, trail significantly in FID (145.22 and 116.48, respectively), underscoring the limitations of standard adversarial approaches in few-shot medical imaging contexts.

Table 1. Comparative performance of generative models on the target PDGM domain. The best is in **bold**, the second best is underlined. ↑/↓ indicate the higher/lower, the better performance. Metrics are reported as mean values aggregated across 3×5-fold cross-validation.

	(a) Fidelity		(b) Downstream Classification		
Model	FID↓	SSIM↑	BAcc↑	F1↑	AUC↑
CGAN [18]	145.22	0.374	0.764	0.720	0.876 ± 0.142
3M-CGAN [19]	116.48	0.680	0.780	0.731	0.866 ± 0.004
VAE-GAN [20]	88.18	**0.750**	0.751	0.675	0.882 ± 0.004
ALDM (K=16, s=0.3)	88.02	0.716	0.780	0.730	0.871 ± 0.004
ALDM (K=16, s=0.5)	88.08	0.714	0.774	0.721	0.877 ± 0.004
ALDM (K=16, s=1.0)	87.52	0.715	0.783	0.733	0.897 ± 0.005
ALDM (K=10, s=3.0)	95.08	0.699	0.856	0.832	0.948 ± 0.003
ALDM (K=16, s=3.0)	**85.40**	0.712	**0.875**	**0.836**	**0.987 ± 0.001**

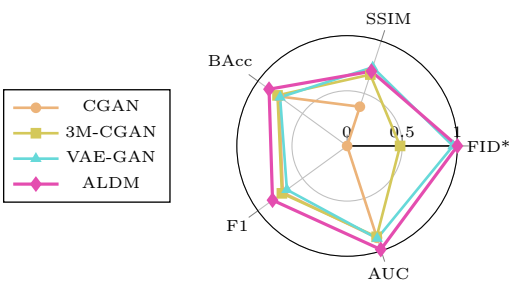

Figure 2. Multi-metric performance comparison (*FID normalized).

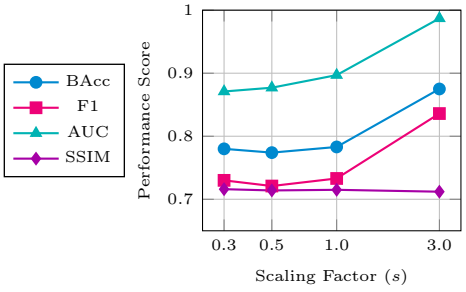

Figure 3. ALDM hyperparameter sensitivity analysis. Performance trends across different scaling factors (s) with fixed K=16.

5.1.2. **Downstream Classification Evaluation**

The clinical utility of the synthesized images was assessed by training a downstream classifier on the real 12k image dataset and testing it on synthetic PDGM samples generated by different models. ALDM ($K = 16, s = 3.0$) outperformed all other models across the primary classification metrics, reaching a balanced accuracy (BAcc) of 0.875, F1 of 0.836, and a superior AUC of 0.987 ± 0.001. In contrast, the VAE-GAN, despite its high SSIM, yielded the lowest F1 score (0.675) and the lowest BAcc (0.751) among the tested models, suggesting that its generated features may not translate as effectively into diagnostic tasks. The proposed model's consistent performance across BAcc, F1, and AUC confirms that the ALDM framework preserves more robust and discriminative features necessary for downstream medical applications.

5.1.3. **Ablation Study**

Impact of Scale Parameter (s). We examine the model's sensitivity to the classifier-free guidance scale parameter (s) at a fixed few-shot size of $K = 16$. As shown in Table 1 and Figure 3, variations in the scale (from $s = 0.3$ to $s = 1.0$) result in relatively subtle performance shifts, with FID scores remaining stable around 88 and AUC values showing consistent improvement. While the default configuration of $s = 3.0$ achieves the optimal balance, marked by a notable peak in AUC (0.987) and the lowest FID (85.40), the model's performance is not overly sensitive to this choice. All tested scales consistently outperform the GAN and VAE-based baselines in downstream classification, indicating that the ALDM framework is robust to the specific tuning of the guidance scale.

Impact of Few-Shot Examples (K). The impact of the number of target-domain images (K) was also evaluated by comparing $K = 10$ and $K = 16$ at a scale of $s = 3.0$. The results indicate that while increasing the sample size to $K = 16$ provides the best overall results in both fidelity and classification accuracy, the difference between the two settings is marginal. Even with only 10 images, the ALDM achieves an AUC of 0.948 and a balanced accuracy of 0.856, which remain superior to those of CGAN, 3M-CGAN, and VAE-GAN. This suggests

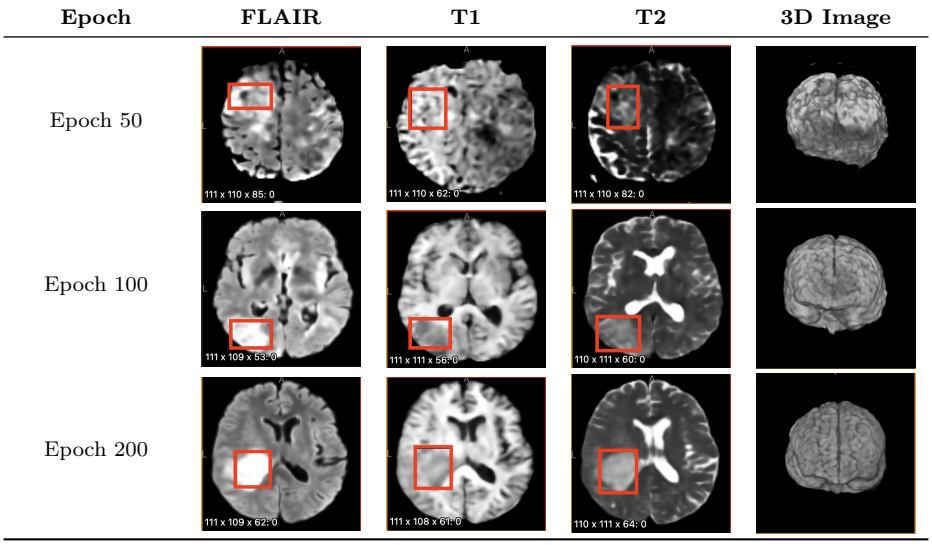

Figure 4. Qualitative evolution of our proposed ALDM with synthesized MRI volumes over training epochs. Progressive improvements in anatomical fidelity, tumor delineation, and cross-modal consistency are observed as training converges.

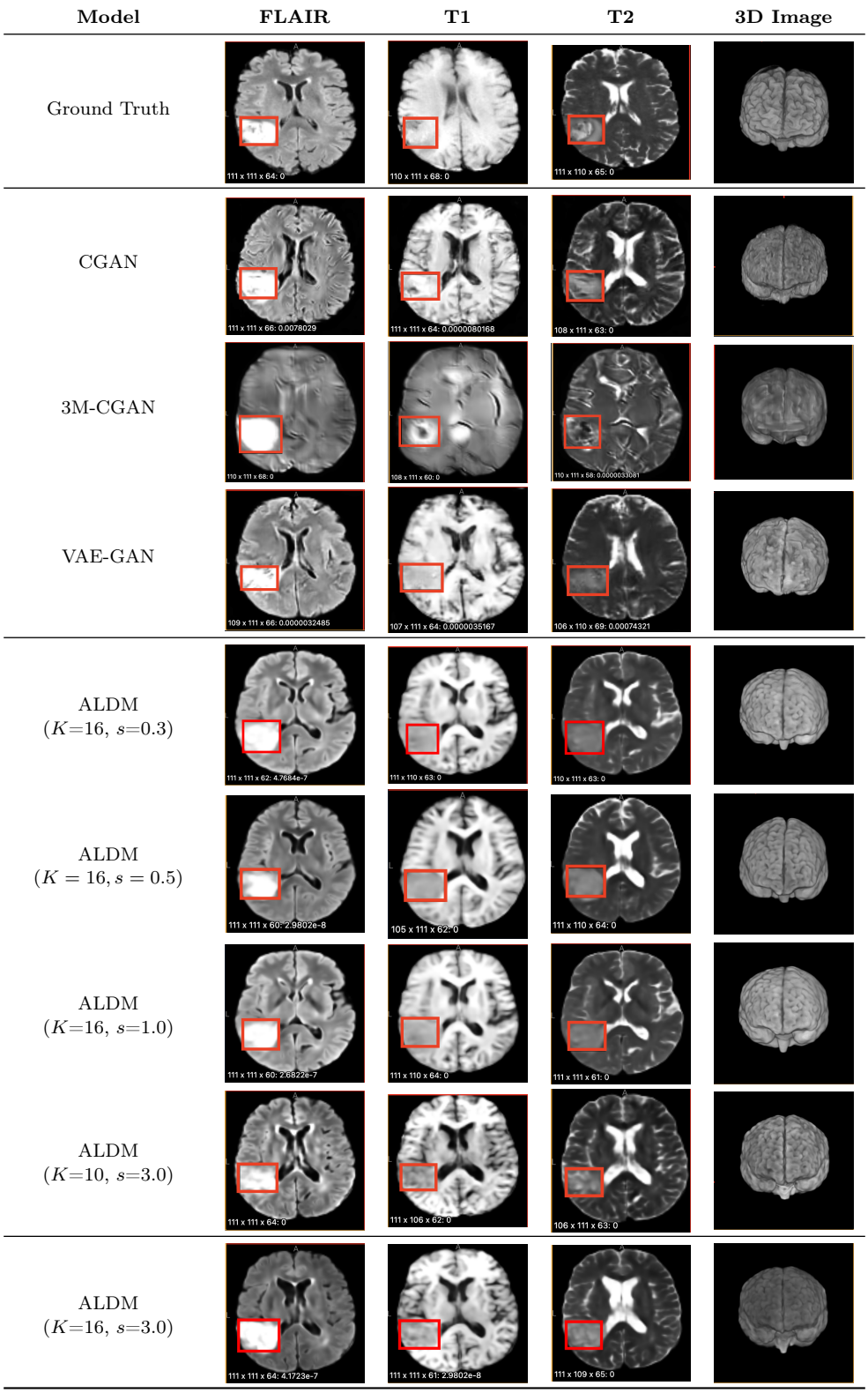

Figure 5. Qualitative comparison of synthesized MRI volumes across modalities. Tumor regions are highlighted in the red box to assess anatomical fidelity, cross-modal consistency, and 3D structural coherence.

that while our chosen default of $K = 16$ is optimal, the method remains highly effective in extremely low-data regimes, demonstrating that the specific choice of K within this range is not critical for achieving competitive results.

This relative stability across different hyperparameter settings suggests that the ALDM framework is inherently robust. While fine-tuning s and K can provide incremental gains in fidelity and downstream accuracy, the overall success of the method is not overly dependent on a specific parameter choice. This flexibility is particularly advantageous in low-resource medical settings where exhaustive hyperparameter optimization may not be feasible.

5.2. **Qualitative Evaluation**

To visually validate the proposed framework's performance, we provide qualitative assessments of the training process and comparisons with baseline methods.

5.2.1. **Progressive Training Performance**

As shown in Figure 4, the ALDM exhibits a clear progression in image quality as training epochs increase. At early stages (epoch 50), the model captures the general brain structure but produces blurry tumor regions with limited modality-specific contrast. By epoch 100, the anatomical fidelity improves significantly, with more distinct boundaries between white and gray matter. At convergence (epoch 200), the model demonstrates sharp tumor delineation and high cross-modal consistency, particularly in the FLAIR and T2 modalities, where pathological features are most prominent.

5.2.2. **Comparative Qualitative Analysis**

Figure 5 illustrates a side-by-side comparison of the synthesized MRI volumes. While the VAE-GAN produces structurally coherent images, they often lack the fine-grained texture and sharp pathology-to-tissue boundaries found in the Ground Truth. The standard GAN baselines (CGAN and 3M-CGAN) exhibit noticeable artifacts and a lack of sharpness in the tumor regions. In contrast, our proposed ALDM ($K = 16, s = 3.0$) generates volumes that most closely resemble the ground truth in terms of both 2D slice detail and 3D structural coherence. Specifically, the tumor regions (highlighted in red) in our model maintain consistent intensity profiles across T1, T2, and FLAIR modalities, whereas other models often struggle with *mode collapse* in the high-intensity FLAIR regions. Furthermore, the ablation variants of ALDM ($s = 0.3$ to $s = 1.0$) exhibit high structural similarity to the default model, confirming that although $s = 3.0$ yields the sharpest results, the framework remains robust across various configurations.

6. **Conclusion**

We proposed the ALDM framework, establishing a benchmark for high-fidelity 3D MRI synthesis in data-scarce, few-shot regimes. By transferring anatomical priors from a data-rich GBM domain into a compact latent space, our method achieves superior distributional similarity and structural consistency while remaining robust to hyperparameter variations. The model's clinical utility is validated by high-performing downstream classifiers, proving it captures essential diagnostic features rather than superficial textures. While demonstrating significant potential for augmenting rare disease data, future work will focus on integrating generative-predictive modules and enhancing pathology-aware control.

Acknowledgment

This work was supported by NSERC Discovery Grant No RGPIN-2025-04478 and NSERC Discovery Supplement Award No DGECR-2025-00129.

References

[1] E. Calabrese et al. "The University of California San Francisco preoperative diffuse glioma MRI dataset". In: *Radiology: Artificial Intelligence* 4.6 (2022), e220058.

[2] S. Bakas et al. "Multi-parametric magnetic resonance imaging (mpMRI) scans for de novo Glioblastoma (GBM) patients from the University of Pennsylvania Health System (UPENN-GBM)". In: *The Cancer Imaging Archive* 10 (2021).

[3] V. Roca et al. "IGUANe: A 3D generalizable CycleGAN for multicenter harmonization of brain MR images". In: *Medical Image Analysis* 99 (2025), p. 103388.

[4] K. Kamnitsas et al. "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks". In: *International conference on information processing in medical imaging*. Springer. 2017, pp. 597–609.

[5] J.-P. Fortin et al. "Harmonization of cortical thickness measurements across scanners and sites". In: *Neuroimage* 167 (2018), pp. 104–120.

[6] M. Bento et al. "Deep learning in large and multi-site structural brain MR imaging datasets". In: *Frontiers in Neuroinformatics* 15 (2022), p. 805669.

[7] C. H. Suh et al. "Imaging prediction of isocitrate dehydrogenase (IDH) mutation in patients with glioma: a systemic review and meta-analysis". In: *European radiology* 29.2 (2019).

[8] Z. Dorjsembe et al. "Conditional diffusion models for semantic 3D brain MRI synthesis". In: *IEEE Journal of Biomedical and Health Informatics* 28.7 (2024), pp. 4084–4093.

[9] N. C. Truong et al. "Synthesizing 3D multicontrast brain tumor MRIs using tumor mask conditioning". In: *Medical Imaging 2024: Imaging Informatics for Healthcare, Research, and Applications*. Vol. 12931. SPIE. 2024, pp. 116–120.

[10] Y. Na et al. "Laplacian filter attention with style transfer GAN for brain tumor MRI imputation". In: *Scientific Reports* 15.1 (2025), p. 35453.

[11] D. Mukherkjee et al. "Brain tumor image generation using an aggregation of GAN models with style transfer". In: *Scientific reports* 12.1 (2022), p. 9141.

[12] B. H. Menze et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[13] S. Bakas et al. "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific data* 4.1 (2017), pp. 1–13.

[14] M. Abbad Andaloussi et al. "Exploring adult glioma through MRI: A review of publicly available datasets to guide efficient image analysis". In: *Neuro-Oncology Advances* 7.1 (2025).

[15] H. G. Pemberton et al. "Multi-class glioma segmentation on real-world data with missing MRI sequences: comparison of three deep learning algorithms". In: *Scientific reports* 13.1 (2023).

[16] M. Zhou et al. "Generating 3D brain tumor regions in MRI using vector-quantization Generative Adversarial Networks". In: *Computers in Biology and Medicine* 185 (2025), p. 109502.

[17] L. Zhang, A. Rao, and M. Agrawala. "Adding Conditional Control to Text-to-Image Diffusion Models". In: *IEEE International Conference on Computer Vision (ICCV)*. 2023.

[18] M. Mirza and S. Osindero. "Conditional generative adversarial nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[19] B. Xin, Y. Hu, Y. Zheng, and H. Liao. "Multi-modality generative adversarial networks with tumor consistency loss for brain mr image synthesis". In: *2020 IEEE 17th international symposium on biomedical imaging (ISBI)*. IEEE. 2020, pp. 1803–1807.

[20] A. B. L. Larsen et al. "Autoencoding beyond pixels using a learned similarity metric". In: *International conference on machine learning*. PMLR. 2016, pp. 1558–1566.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

Appendix A. User Interface

To facilitate qualitative evaluation and interactive exploration of the proposed framework, we developed a web-based user interface that enables on-demand generation of synthetic MRI volumes for both the GBM and PDGM domains[1]. As illustrated in Figure 6, the interface allows users to directly compare generated samples against their corresponding ground-truth images, providing intuitive insight into anatomical fidelity, tumor morphology, and cross-domain behavior.

Given the emphasis on few-shot image generation, the interface is designed to highlight both controlled and unconstrained synthesis. Specifically, the first 16 generated samples reuse tumor masks aligned with the ground-truth images, resulting in synthetic volumes with tumors located at identical spatial positions. This setting isolates the model's ability to preserve anatomical structure and intensity characteristics under fixed spatial constraints. All subsequent generated samples reuse the same set of masks but apply them to different anatomical contexts, producing tumors at novel spatial locations and enabling assessment of the model's capacity for spatial generalization beyond the observed target-domain examples.

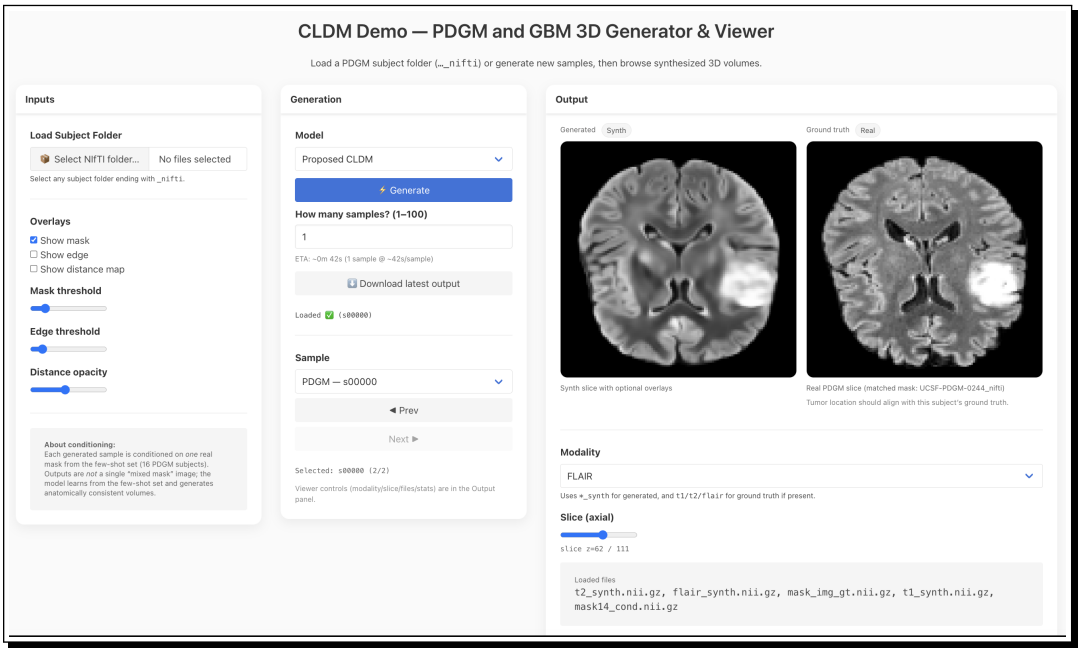

Figure 6. Web-based user interface for interactive MRI generation and qualitative comparison between ground-truth and synthetic volumes.

[1]A demonstration of the user interface is available at https://youtu.be/OARkHCNygmY.