Conditional Latent Diffusion Model for Synthetic 3D Brain MRI Generation to Enhance Temporal Lobe Epilepsy Detection

Jay Sawant
Halıcıoğlu Data Science Institute
University of California San Diego
La Jolla, CA 92093
jsawant@ucsd.edu

Abstract

Temporal Lobe Epilepsy (TLE) detection relies on large annotated datasets of brain MRI scans that are often scarce and heterogeneous. To address these challenges, we propose a novel pipeline that combines Variational Autoencoder–Generative Adversarial Network (VAE-GAN) compression with a conditional diffusion model for generating high-fidelity 3D brain MRIs. First, our VAE-GAN reduces the 3D volumes to a low-dimensional latent representation while retaining key structural features. Next, a conditional latent diffusion model is trained to iteratively refine noise samples into latent codes corresponding to healthy controls or TLE patients, enabling class-specific generation. Finally, the decoder reconstructs high-resolution 3D MRI scans from these latent codes, producing anatomically realistic volumes. By augmenting training data with these synthetic scans, we aim to substantially improve TLE detection accuracy in scenarios with limited labeled data, thereby leveraging the advantages of diffusion-based synthesis for clinical applications.

1 Introduction

Temporal Lobe Epilepsy (TLE) is a complex neurological disorder characterized by subtle structural anomalies—especially in the hippocampus—that challenge accurate diagnosis. The reliance on high-resolution, volumetric brain MRI scans for TLE detection is hindered by the scarcity and heterogeneity of large, labeled datasets due to privacy concerns, high acquisition costs, and inter-institutional variability. To mitigate these challenges, generative modeling offers a promising route to augment existing datasets with synthetic yet clinically meaningful data. In this work, we propose a conditional latent diffusion framework that integrates a Variational Autoencoder–Generative Adversarial Network (VAE-GAN) for effective latent space compression with a 3D Diffusion Transformer (DiT-3D) for sample generation. By generating condition-specific synthetic 3D MRIs that capture fine-grained pathological details, our approach aims to enhance TLE detection accuracy through improved training data diversity.

2 Background & Related Work

Recent advances in generative modeling for medical imaging have shown promise in addressing issues of data scarcity, privacy concerns, and the high costs of clinical annotation. In particular, diffusion models have gained traction due to their ability to generate high-fidelity samples and offer fine control over the synthesis process. For instance, Med-DDPM—a 3D semantic diffusion model for brain MRI synthesis—demonstrated that incorporating semantic conditioning via channel-wise

Preprint. Under review.

concatenation of relevant anatomical structures yields images that are both anatomically coherent and visually realistic [1]. Notably, Med-DDPM achieved competitive Dice scores in tumor segmentation tasks, suggesting that diffusion-based synthetic data can enhance downstream clinical performance.

Another significant contribution is the MAISI framework (Medical AI for Synthetic Imaging) [2], which generates high-resolution 3D CT images using a latent diffusion model built upon a VAE-GAN volume compression network [5]. This pipeline offers flexibility in handling various volume dimensions and voxel spacings, which is critical for capturing diverse anatomical variations. By incorporating additional conditioning inputs through ControlNet, MAISI accurately produces annotated synthetic volumes suitable for organ segmentation tasks, thereby underscoring the versatility of diffusion-based methods across imaging modalities. Our work leverages MAISI's VAE-GAN component to encode 3D brain MRIs into a compact latent space.

On the classification front, 3D convolutional neural networks (CNNs) have been extensively utilized for Temporal Lobe Epilepsy (TLE) diagnosis [4]. These models emphasize the importance of 3D feature extraction in capturing subtle hippocampal anomalies that may be missed by traditional 2D CNNs, yet they require large volumes of labeled data and high amount of compute. In this work, we stick with the 2D CNN methods due to compute limitations. We propose augmenting 2D CNN pipelines with synthetic MRI scans that incorporate subtle TLE indicators.

Given the nuanced nature of TLE pathology—particularly the minor intensity and shape variations in the hippocampal region—we advocate for a 3D Diffusion Transformer based on DiT-3D [6] for volume generation, as an alternative to conventional UNet-based architectures. Transformers, with their self-attention mechanisms, effectively incorporate global context, thereby enhancing the generator's capacity to capture fine-grained variations. By integrating the VAE-GAN strategy from MAISI with a 3D Diffusion Transformer, our approach bridges the gap between robust volume compression and precise modeling of subtle pathological details, ultimately bolstering the efficacy of TLE detection models without over-reliance on extensive labeled datasets.

3 Methods

3.1 VAE-GAN Architecture for Latent Volume Compression

Our first step is to compress high-resolution 3D brain MRIs into a lower-dimensional latent space while preserving essential structural information. To accomplish this, we adopt a Variational Autoencoder–Generative Adversarial Network (VAE-GAN) architecture, as illustrated in Figure 1. The VAE portion comprises an **encoder** and a **decoder**, with the encoder mapping input volumes to a mean (μ) and standard deviation (σ) in a latent space, and the decoder reconstructing the corresponding 3D brain MRI from latent codes. This approach serves two main purposes: (1) reducing the dimensionality of the data, thereby making downstream diffusion-based generation more tractable, and (2) enabling smooth latent manipulations that aid in conditional generation tasks.

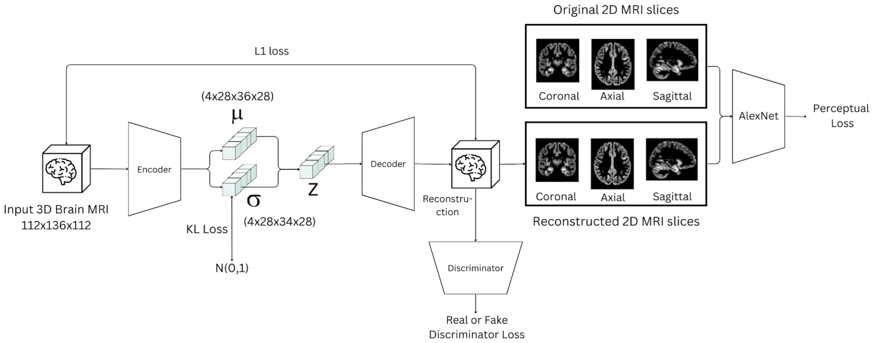

Figure 1: Overview of the VAE-GAN architecture for compressing and reconstructing 3D brain MRIs. The encoder outputs mean (μ) and standard deviation (σ) feature maps, forming a latent distribution from which the latent code z is sampled. The decoder then reconstructs the 3D volume, which is passed to the discriminator for a real/fake decision. Additionally, a perceptual loss is computed based on 2D slices fed into a pretrained network of AlexNet.

Latent Sampling & KL Loss. As in a standard Variational Autoencoder, a latent code z is sampled from $\mathcal{N}(\mu, \sigma)$, enforcing a prior distribution of $\mathcal{N}(0, 1)$. The Kullback–Leibler (KL) divergence term penalizes deviations of the learned distribution from the prior, promoting smoother latent representations and regularization:

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}\Big[\mathcal{N}(\mu, \sigma) \,\big\|\, \mathcal{N}(0, 1)\Big]. \tag{1}$$

Discriminator. To enhance the perceptual quality of reconstructions, we incorporate an adversarial discriminator that classifies each reconstructed volume (or its 2D slices) as either *real* or *fake*. This adversarial loss incentivizes the decoder to produce detailed, realistic outputs:

$$\mathcal{L}_{\mathrm{adv}} = \mathbb{E}\big[\log D(x)\big] + \mathbb{E}\big[\log(1 - D(\hat{x}))\big], \tag{2}$$

where x is a real MRI volume and $\hat{x}$ is the reconstructed (or generated) one.

Reconstruction & Perceptual Loss. We employ an ℓ_1 reconstruction loss between the input and decoded volumes to encourage voxel-level fidelity instead of mean-squared loss as it is sensitive to outliers.

$$\mathcal{L}_{\ell_1} = \|x - \hat{x}\|_1. \tag{3}$$

In addition, a perceptual loss is computed by comparing feature activations of original and reconstructed 2D slices through a pretrained network of AlexNet. This loss captures higher-level semantic information, leading to more anatomically plausible reconstructions:

$$\mathcal{L}_{\mathrm{perc}} = \sum_i \|\phi_i(x) - \phi_i(\hat{x})\|_2, \tag{4}$$

where ϕ_i denotes the activations from the i-th layer of the pretrained model.

Combined Objective. Overall, the VAE-GAN is trained to minimize a combination of the KL divergence, reconstruction (ℓ_1), perceptual, and adversarial losses:

$$\mathcal{L}_{\mathrm{VAE-GAN}} = \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} \;+\; \lambda_{\ell_1}\mathcal{L}_{\ell_1} \;+\; \lambda_{\mathrm{perc}}\mathcal{L}_{\mathrm{perc}} \;+\; \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}} \tag{5}$$

By balancing these terms, the VAE-GAN effectively captures essential structural features in the latent domain while preserving the detailed anatomical information necessary for downstream tasks.

3.2 Conditional Diffusion Model using DiT3D

In our framework, we adopt a conditional diffusion model based on DiT3D [6], originally designed for 3D shape generation. The core idea of diffusion models is to learn the reverse process of a gradually noisy data generation procedure. In our adaptation, we modify the DiT3D block by removing the voxelization and de-voxelization steps, thereby allowing the diffusion process to operate directly in the latent feature space.

Diffusion Process. Let z_0 denote the latent feature representing a 3D brain MRI. The forward diffusion process progressively adds Gaussian noise over T timesteps:

$$q(z_t \mid z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1 - \beta_t}\, z_{t-1}, \beta_t\, \mathbf{I}\right), \tag{6}$$

where β_t is a variance schedule controlling the noise magnitude at each timestep t. The reverse (denoising) process is parameterized by a transformer-based network, which learns to estimate the original latent feature z_0 from a noisy observation z_t.

Conditional Generation. To steer the diffusion process towards generating latent features that are consistent with desired characteristics of the abnormality y, we condition the reverse diffusion process as follows:

$$p_\theta(z_{t-1} \mid z_t, y) = \mathcal{N}\big(z_{t-1};\, \mu_\theta(z_t, t, y),\, \Sigma_t\big),$$

where $\mu_\theta(\cdot)$ is a learnable function implemented via the DiT3D transformer architecture, and Σ_t is a (fixed) variance schedule that does not depend on θ. This conditional setup ensures that the generated latent features $\{z\}$ are aligned with the intended properties of brain MRIs.

Integration with VAE-GAN Decoder. The latent features z produced by the diffusion process serve as inputs to a VAE-GAN decoder. This decoder reconstructs high-fidelity 3D brain MRIs from the latent space. By decoupling the latent feature generation (handled by the diffusion model) from the image synthesis (performed by the VAE-GAN decoder), the system leverages the strengths of both models: the diffusion model efficiently captures global structural information, while the VAE-GAN refines local details and textures.

Modifications to DiT3D. A key modification in our approach is the removal of the voxelization and de-voxelization operations present in the original DiT3D design.

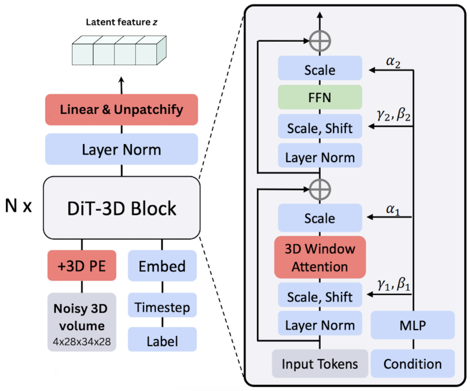

Figure 2: Illustration of DiT3D block. This block is adapted from the DiT-3D paper[6] for 3D shape generation. The diffusion transformer takes a 3D noisy latent **z** as input, and a patchification operator is used to generate token-level patch embeddings, where 3D positional embeddings are added. 3D window attention is addded to restrict attention to nearby spatial regions, so that the model learns local dependencies in the 3D data. Finally, after stacking N such blocks, the unpatchified voxel tensor output from a linear layer is used to predict the noise in the input noisy feature.

4 Experimental Setup

Dataset: We conduct our experiments on T1-weighted MRI scans from 3540 participants: 2110 diagnosed with temporal lobe epilepsy (TLE) and 1430 healthy controls (HCs). The data were gathered from twelve sites: the Medical University of South Carolina, Emory University, New York University, Northwell University, University of Bonn, University of Pittsburgh, University of Pennsylvania, Rush University, University of California San Diego, University of California San Francisco, the University of Liverpool, and the Human Connectome Project [4]. All TLE patients fulfilled the criteria for drug-resistant unilateral TLE, verified by clinical, neurophysiological, and radiographic assessments, and had no extratemporal lesions or alternative neurological disorders. Institutional Review Board approvals were granted at each institution, and informed consent was obtained following ethical guidelines.

Preprocessing: We used the *Nii_preprocess* toolkit in conjunction with SPM12 and CAT12. Preprocessing involved spatial normalization to the MNI152 template space (resampled to $112 \times 136 \times 112$ voxels), tissue segmentation, and smoothing at 10 mm full-width at half-maximum (FWHM). Labeling was subsequently applied to facilitate anatomical analysis. These preprocessing steps aim to mitigate inter-individual anatomical variability while retaining crucial structural details.

Evaluation Metrics. To assess the effectiveness of the proposed conditional diffusion model and the resulting 3D brain MRI generations, we employ several complementary evaluation metrics. We measure a perceptual loss on a held-out validation set by comparing high-level feature representations (extracted from a pretrained AlexNet which was trained against ImageNet) of the generated MRI volumes against their corresponding ground truth or reference volumes. This metric reflects the fidelity of structural details captured in the synthetic data. Reconstruction loss is used to quantify

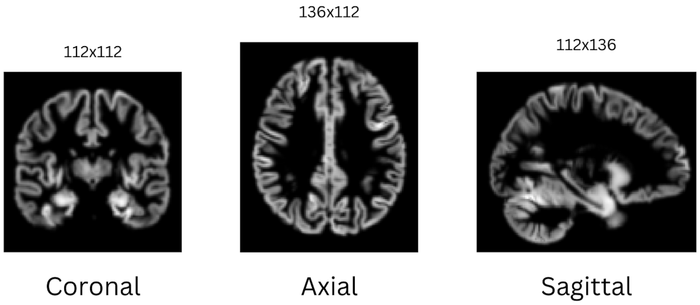

Figure 3: Example T1-weighted MRI slices from the dataset, all taken at the 70th slice in each orientation. From left to right: Coronal view (112×112), Axial view (136×112), and Sagittal view (112×136).

voxel-wise differences between synthetic and real MRI volumes. A lower reconstruction loss indicates more accurate synthesis of brain structures and intensities.

Fréchet Inception Distance (FID). Although primarily introduced for natural image generation, the FID score is also adopted here as a sample-level measure of similarity between real and generated MRI distributions[3]. Specifically, we embed real and synthetic 2D slices from 3D volumes into a feature space (using pre-trained Inception V3 network[7]) and compute the Fréchet distance between their Gaussian approximations. A lower FID indicates that the synthetic dataset is more similar to the real dataset.

Qualitative Expert Evaluation In addition to the quantitative metrics described above, we will conduct a qualitative evaluation of the generated 3D brain MRI volumes by domain experts. Specifically, we will select a set of generated samples from our diffusion model and pair them with real MRI volumes. We will then present these unlabeled scans to a neuroscientist, who will attempt to classify each volume as "real" or "synthetic". This setup serves as a "Turing test" for 3D MRI realism and also allows us to gauge expert perception of subtle artifacts or inconsistencies that might not be captured by automated metrics.

5 Implementation and Results

5.1 VAE-GAN Training Setup

In this work, we first train a VAE-GAN to compress and reconstruct 3D brain MRI scans. We use an 80%–20% split for training and validation, selecting the final model checkpoint based on the lowest validation loss. This validation loss is a weighted sum of the reconstruction (ℓ_1), perceptual (LPIPS), Kullback–Leibler (KL), and adversarial losses, as described in Section 3.

Network Architecture. The VAE encoder contains three ResNet-style blocks of which the first two performs downsampling, leading to a compressed latent representation. The final compression factor is 16x (112x136x112 being compressed to 4x28x34x28). We apply Group Normalization[9] within each block to facilitate stable training given our small batch size of 2. The decoder mirrors the encoder structure with corresponding ResNet blocks in reverse, progressively upsampling the latent features back to the original spatial dimensions.

Training Details. We initially train the VAE for 100 epochs using only the L1 reconstruction loss, LPIPS loss, and KL divergence loss. This choice is driven by limited computational resources (we use an Nvidia RTX 4090 GPU with 24GB of memory). We employ the Adam optimizer with an initial learning rate of 1×10^{-3}, along with a cosine annealing with warm restarts scheduler to periodically reduce the learning rate.

After this first training phase, we switch to a two-script setup to train the discriminator and the VAE separately. We first train the discriminator for 2 epochs, then fine-tune the VAE by including the adversarial loss term (in addition to L1, LPIPS, and KL losses). We stop VAE training whenever the running average of the adversarial loss falls below 0.1. We then update the discriminator again using the newly fine-tuned VAE checkpoint. This iterative procedure of alternating between VAE and discriminator training is repeated for 65 iterations until we achieve the most visually appealing results.

Throughout the training process, we use LPIPS computed via a pretrained AlexNet on ImageNet for the perceptual component. The following weighting coefficients strike a balance between anatomical fidelity and perceptual realism, while keeping the latent space smooth and disentangled:

- Reconstruction loss (L_{ℓ_1}): weight = 1.0
- Perceptual (LPIPS) loss (L_{perc}): weight = 0.6
- KL divergence (L_{KL}): weight = 1×10^{-6}
- Adversarial loss (L_{adv}): weight = 0.02

We save model checkpoints at each epoch or iteration as appropriate and select the best generator based on the combined loss metric on the validation set.

VAE-GAN Results. Table 1 summarizes our reconstruction metrics on the validation set. They indicate that the VAE-GAN is able to learn a meaningful latent representation of the 3D MRI volumes, with relatively low reconstruction and perceptual losses. Visual inspection of the reconstructions suggests anatomically plausible outputs.

Table 1: VAE-GAN validation results on reconstructed 2D slices. We report L_1 error, LPIPS, and FID. Although FID is typically used to measure the distribution similarity between two distinct image sets, here it is calculated between original and reconstructed slices, so we expect a value close to zero.

	L_{ℓ_1}	LPIPS	FID
VAE-GAN	0.058	0.017	0.33

5.2 Conditional Latent Diffusion Model

Following the VAE-GAN training, we train a conditional latent diffusion model inspired by DiT-3D [6]. This diffusion model operates on the VAE's latent space (of size (`4,28,34,28`)) rather than on high-dimensional 3D MRI volumes. By doing so, we benefit from more efficient learning and generation while still capturing the relevant anatomical structures.

Architecture and Conditioning. We use a Transformer-based backbone with 12 layers (Depth N), a hidden size of 384 (d_{emb}), a patch size of $(4, 2, 4)$, and 6 attention heads. To enable conditional generation, we introduce a set of learnable class embeddings corresponding to "healthy" and "TLE" categories, plus an additional learnable "class dropout" embedding that the model can attend to when no specific class condition is applied. These embeddings are appended to the patch tokens at each diffusion timestep, guiding the denoising process toward the desired output class.

Noise Scheduler and Loss. We employ a linear noise scheduler over $T = 1000$ timesteps, with β values increasing from $\beta_{\text{start}} = 1 \times 10^{-4}$ to $\beta_{\text{end}} = 2 \times 10^{-2}$. At each timestep, the model predicts the noise $\boldsymbol{\epsilon}_\theta$ added to the noisy latent $\mathbf{z}_t$. Training minimizes the mean squared error (MSE) between the predicted noise and the true noise:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{z}_0, t, \boldsymbol{\epsilon}} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, y) \right\|^2, \tag{7}$$

where $y \in \{\text{healthy, TLE, dropout}\}$ is the conditioning label (or embedding).

Training Configuration. We train the diffusion model for 10,000 epochs using a batch size of 16 on an NVIDIA RTX 4090 (24 GB). The Adam optimizer is used with a fixed learning rate of

$\text{lr} = 1 \times 10^{-4}$. At test time, class-conditional samples are generated by sampling random Gaussian noise $\mathbf{z}_T$ and iteratively denoising it with the learned reverse diffusion process. The final latent samples $\mathbf{z}_0$ are then passed through the pretrained VAE-GAN decoder to reconstruct high-resolution 3D brain MRI scans.

Results. We evaluated the realism of our generated samples through both quantitative and qualitative analyses. Table 2 summarizes the Fréchet Inception Distance (FID) scores computed using an Inception V3-based feature extractor. Healthy synthetic volumes (versus real counterparts) had a FID of 87.36, while TLE synthetic volumes (versus real counterparts) had a FID of 99.07. Although these values are considerably higher than typical FIDs reported on natural images, this reflects the unique challenges of comparing 3D medical volumes in a 2D feature space.

Table 2: Quantitative analysis using FID scores (Inception V3). For each dataset, we extract all 2D slices from the 3D volumes and compare real vs. synthetic distributions.

Dataset	**FID Score**
Healthy (Real vs Synthetic)	87.36
TLE (Real vs Synthetic)	99.07

For the qualitative evaluation, we performed a “Turing test” with a neuroscientist, as shown in Table 3. Out of 48 shuffled 3D scans (24 real and 24 synthetic), only 7 synthetic volumes were correctly identified, corresponding to a detection accuracy of 29.1%. This implies that generated scans are generally difficult to distinguish from real data. We also tested a CNN (pretrained only on real data) on classifying 24 synthetic TLE/Healthy scans. The model correctly classified 19 out of 24, yielding an accuracy of 79.1%, close to its baseline performance of 81.3% when evaluated on real data.

Table 3: Qualitative analysis results. The neuroscientist was presented with shuffled 3D scans (real and synthetic) for detection, and a pretrained CNN was used to classify additional synthetic scans as TLE vs. Healthy.

Dataset	**Synthetics Detection**	**Pretrained CNN**	**Accuracy**
48 shuffled 3D scans (24 real + 24 synthetic)	7/24 correctly detected	N/A	29.1%
24 synthetic scans (balanced data)	N/A	19/24 correctly classified	79.1%

These findings demonstrate that the synthetic volumes bear close similarity to real 3D MRI scans, both visually and in feature space. Although there is still a gap in quantitative metrics like the FID, the qualitative assessment and CNN-based classification suggest that our approach produces sufficiently realistic synthetic 3D volumes to be useful for data augmentation and diagnostic modeling.

5.3 Classifier Design and Evaluation

5.3.1 Network and Training Protocol

We use the EfficientNet-V2[8] architecture for binary TLE vs. healthy classification, using coronal view of the 3D volume. The first convolution layer is adapted to accept 136 input channels (corresponding to the coronal view). A sigmoid output layer performs the final binary decision.

All experiments use the same 80%–20% train-test split of the real MRI scans as the one which was used to train the VAE-GAN and diffusion model. From the 20% test split, we further set aside a validation subset (37%) and use rest of the 300 scans as the final test dataset (63%). We train for 100 epochs with batch size 128, using the Adam optimizer with an initial learning rate of 1×10^{-3} and a cosine annealing with warm restarts schedule. Model performance is evaluated via Accuracy, Precision, Recall, F1-score, AUC-ROC, and AUCPR.

5.3.2 Results With and Without Synthetic Data

Table 4 summarizes the classification metrics under four different training configurations.

Table 4: Binary classification results using EfficientNet-V2 on coronal slices. We report F1-score, Precision, Recall, Accuracy, AUC-ROC, and AUCPR.

Train Dataset	**F1**	**Precision**	**Recall**	**Accuracy**	**AUC-ROC**	**AUCPR**
Real only	0.826	0.886	0.773	81.3%	0.887	0.925
Synthetic only	0.772	0.847	0.710	76.0%	0.842	0.888
Real + Synthetic (1:0.25)	0.845	**0.93**	0.773	**83.67**%	**0.90**	**0.94**
Real + Synthetic (1:0.5)	**0.848**	0.907	**0.796**	83.67%	0.89	0.925

Overall, these results underscore the value of synthetic data in augmenting real MRI scans for TLE detection. Although using only synthetic samples underperforms the real-data baseline, combining real and synthetic volumes significantly enhances model performance. Specifically, adding synthetic data at a 1:0.25 ratio boosts precision (from 0.886 to 0.93) and increases accuracy (from 81.3% to 83.67%), while a 1:0.5 ratio yields a higher recall of 0.796 at the same accuracy. These findings indicate that carefully balanced synthetic augmentation can strengthen both precision and recall, highlighting synthetic MRIs as a promising tool for improving classifier robustness in limited-data scenarios.

6 Conclusion

In this work, we introduced a novel framework for synthesizing three-dimensional (3D) brain MRI scans aimed at enhancing the detection of Temporal Lobe Epilepsy (TLE). The proposed pipeline combines a VAE-GAN for latent compression with a conditional latent diffusion model based on a 3D Diffusion Transformer backbone (DiT-3D). Through extensive experiments, we demonstrated that compressing high-resolution 3D MRI volumes into a latent space not only reduced computational overhead but also preserved essential anatomical features relevant to TLE pathology. Our conditional diffusion approach enabled the generation of class-specific latent codes for both healthy and TLE groups, which, once decoded, exhibited high fidelity and anatomical plausibility. Quantitative measures such as Fréchet Inception Distance (FID) indicated a reasonable alignment between real and synthetic data distributions, whereas a “Turing test” with a domain expert suggested that many synthetic volumes were challenging to distinguish from real scans.

Importantly, our experiments on a TLE-versus-healthy classification task indicated that augmenting real data with synthetic scans improved model robustness and overall diagnostic performance. Specifically, balancing the addition of synthetic 3D MRIs (e.g., at a ratio of 1:0.25 or 1:0.5 relative to real data) enhanced precision and recall across multiple metrics, underlining the value of synthetic data in mitigating data scarcity issues. Furthermore, the classifier trained on synthetic plus real data preserved strong discriminative power for subtle anomalies characteristic of TLE.

Future extensions of this work could explore additional conditioning signals—such as detailed morphological annotations or multi-contrast sequences (e.g., T2-weighted or FLAIR)—to further refine the granularity of generated anomalies. Moreover, integrating domain adaptation strategies may help align synthetic and real distributions across diverse clinical sites and imaging protocols. Overall, our findings underscore the promise of diffusion-based generative models in creating realistic 3D medical images, thereby contributing to both improved diagnostic modeling and the broader pursuit of data-efficient deep learning in clinical neuroimaging.

References

[1] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, July 2024.

[2] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. Maisi: Medical ai for synthetic imaging, 2024.

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[4] Erik Kaestner, Reihaneh Hassanzadeh, Ezequiel Gleichgerrcht, Kyle Hasenstab, Rebecca W Roth, Allen Chang, Theodor Rüber, Kathryn A Davis, Patricia Dugan, Ruben Kuzniecky, Julius Fridriksson, Alexandra Parashos, Anto I Bagić, Daniel L Drane, Simon S Keller, Vince D Calhoun, Anees Abrol, Leonardo Bonilha, and Carrie R McDonald. Adding the third dimension: 3d convolutional neural network diagnosis of temporal lobe epilepsy. *Brain Communications*, 6(5):fcae346, 10 2024.

[5] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric, 2016.

[6] Shentong Mo, Enze Xie, Ruihang Chu, Lewei Yao, Lanqing Hong, Matthias Nießner, and Zhenguo Li. Dit-3d: Exploring plain diffusion transformers for 3d shape generation, 2023.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015.

[8] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021.

[9] Yuxin Wu and Kaiming He. Group normalization, 2018.

A Appendix

A.1 VAE-GAN Reconstructions and Diffusion synthesis

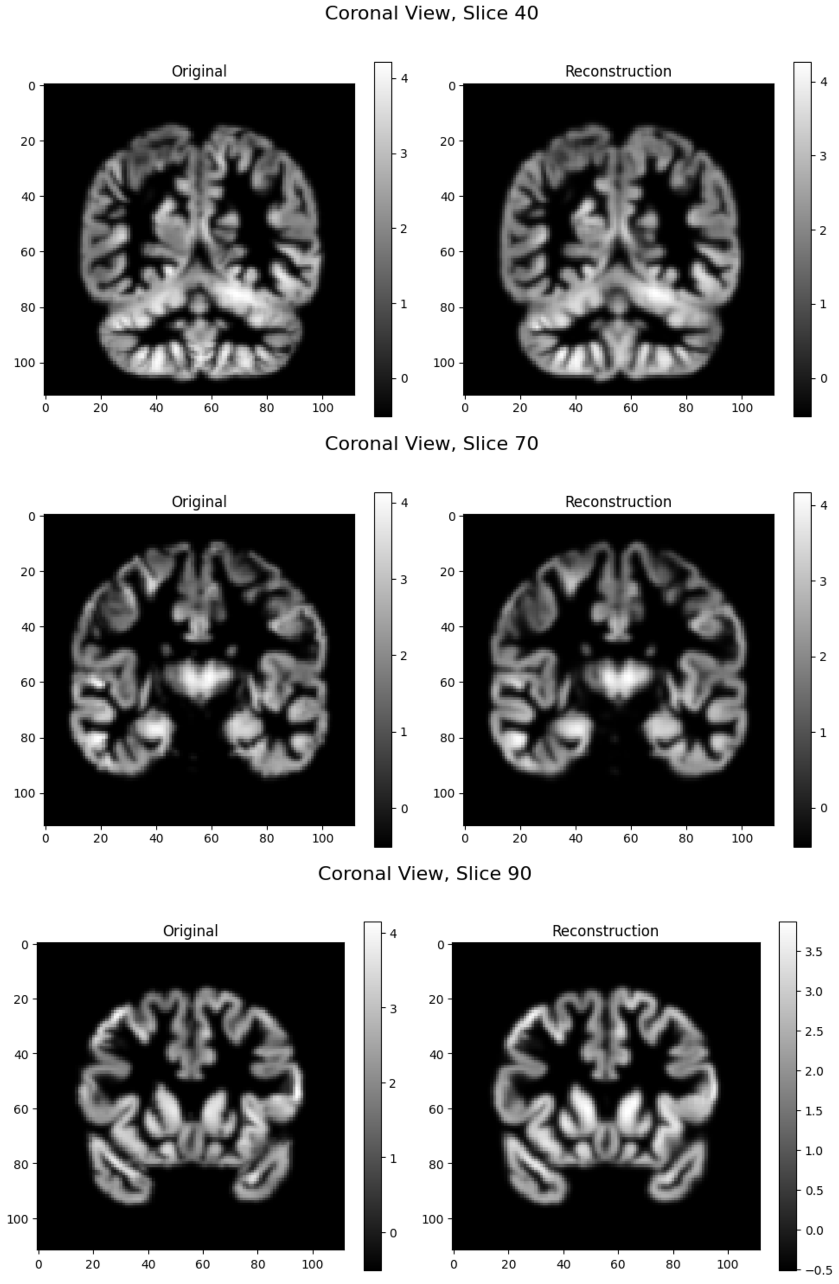

Figure 4: VAE-GAN reconstructions. Example slices from coronal view comparing original MRI (left) and VAE output (right).

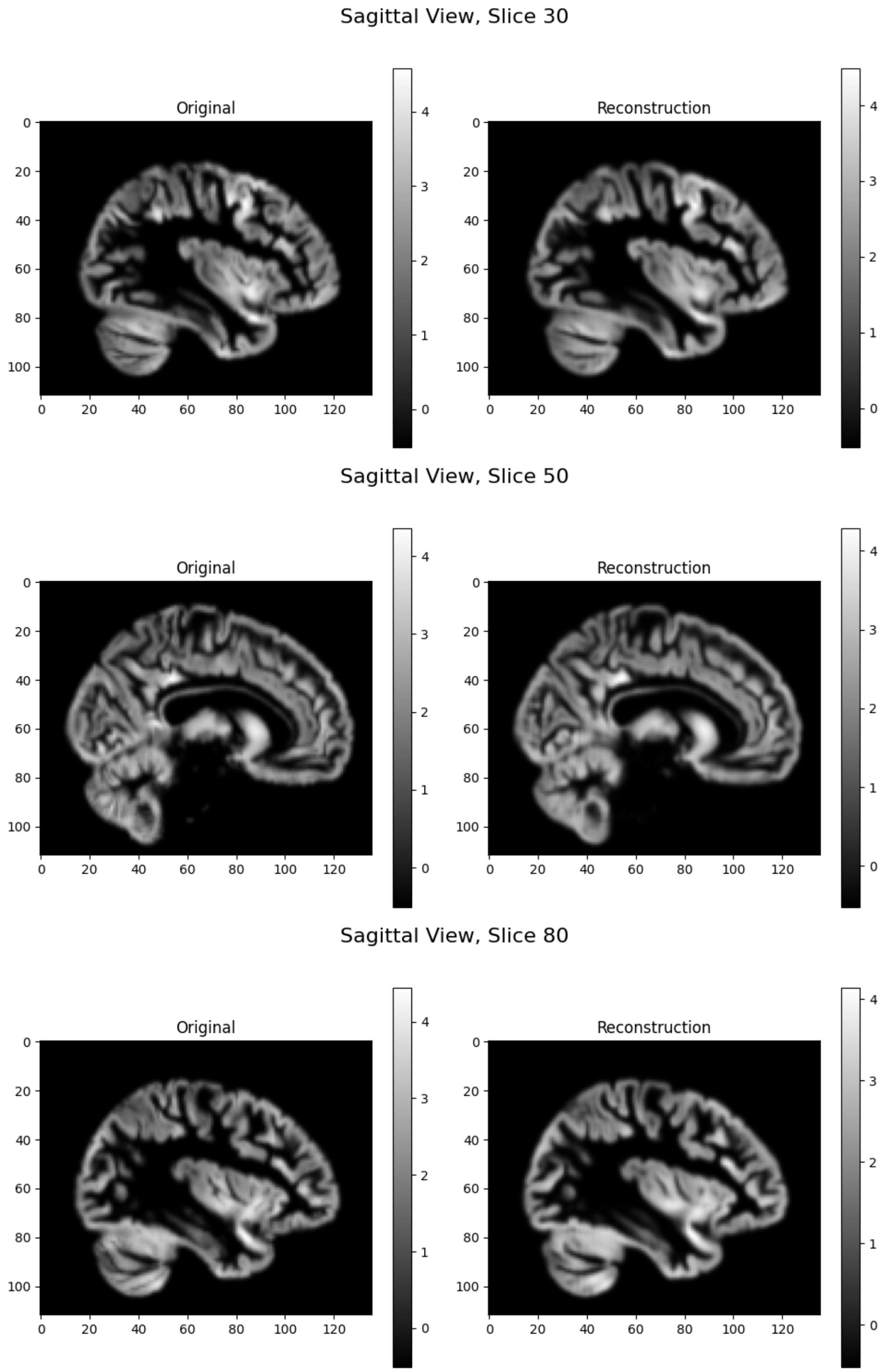

Figure 5: VAE-GAN reconstructions. Example slices from sagittal view comparing original MRI (left) and VAE output (right).

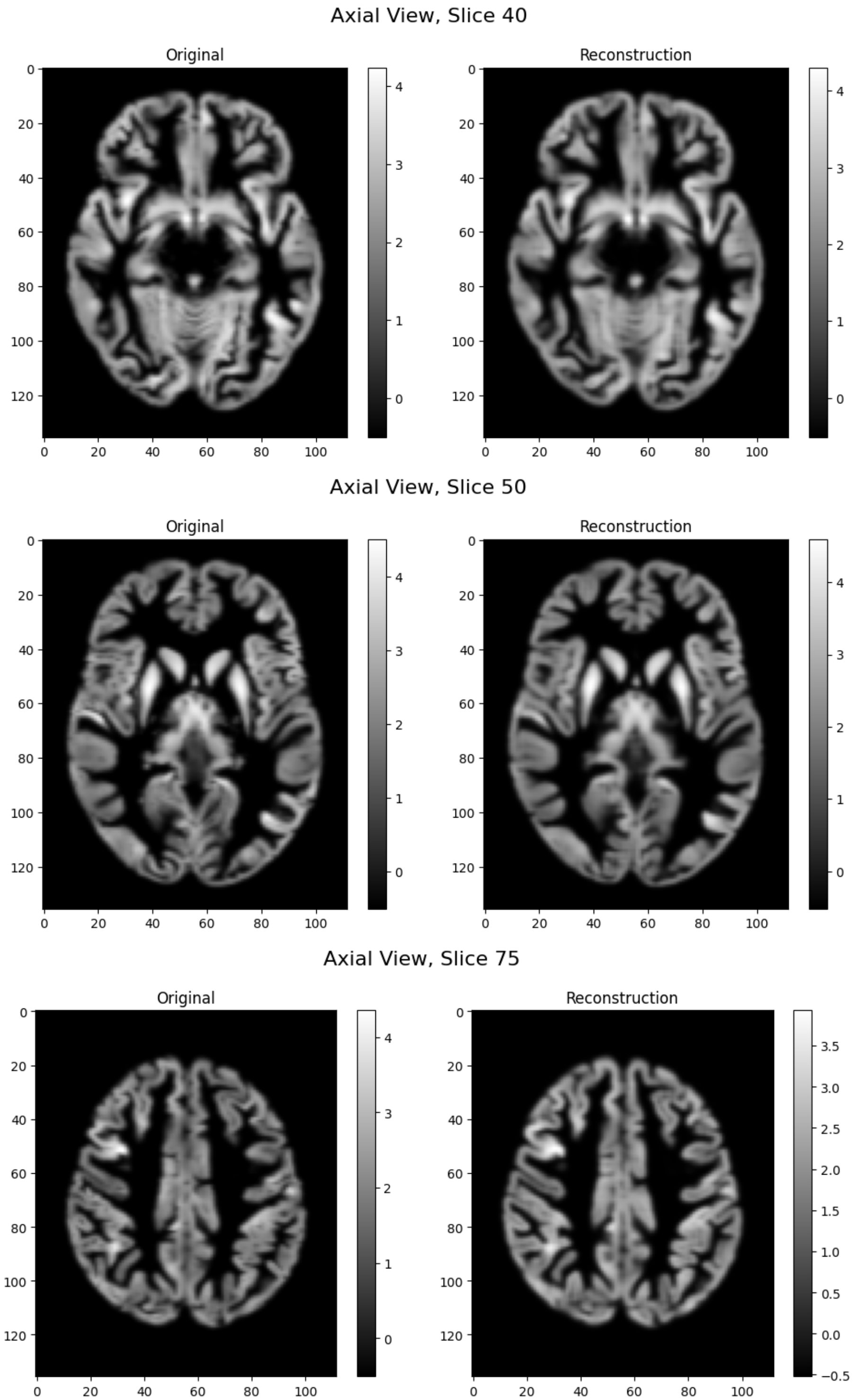

Figure 6: VAE-GAN reconstructions. Example slices from axial view comparing original MRI (left) and VAE output (right).

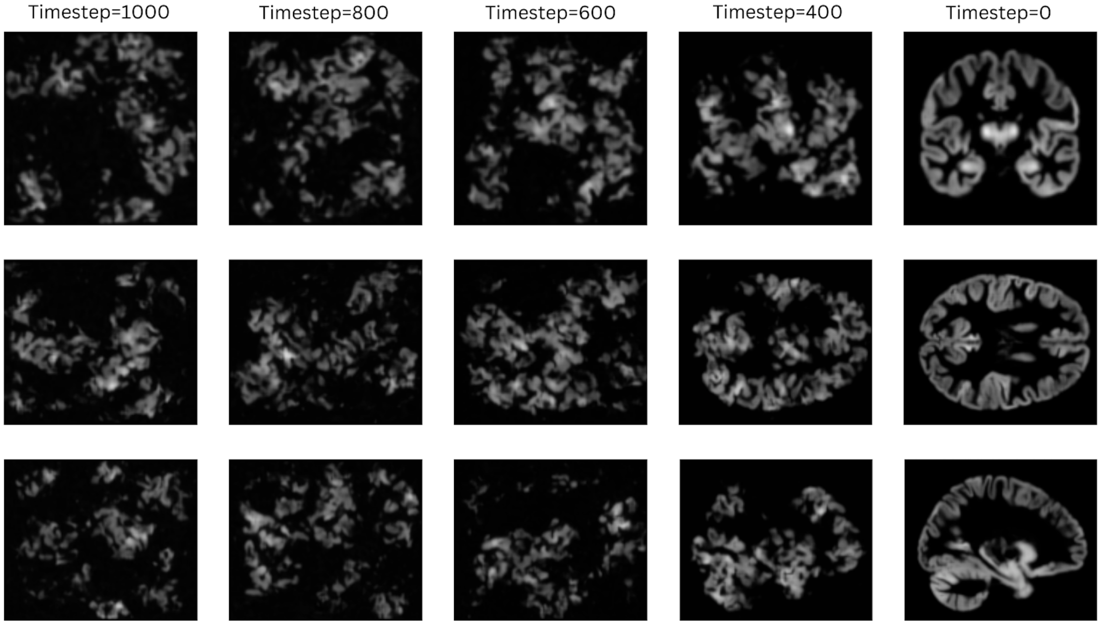

Figure 7: Illustration of the reverse diffusion process for 3D MRI synthesis at selected timesteps. At timestep $= 1000$, the latent representation is heavily corrupted by noise. As the diffusion model iteratively denoises the sample (moving from left to right), finer anatomical details emerge. By timestep $= 0$, the generated slice closely resembles a realistic T1-weighted MRI view.