3D Brain MRI Generation with a Clinically-Conditioned VAE-GAN and Diffusion-Driven Feature Sampling

Najmeh Mashhadi, Emmanouil Nikolakakis, Razvan Marinescu*
Department of Computer Science
University of California, Santa Cruz

Abstract

We introduce a 3D VAE-GAN framework that synthesizes brain MRI volumes conditioned on seven clinical attributes, such as Alzheimer's disease (AD) diagnostic labels and key volumetric measures, including the hippocampus, amygdala, and lateral ventricle, which are known to correlate with AD. Leveraging a 3D encoder-decoder with depthwise-separable convolutions and a style-based modulation, our model efficiently captures critical biomarkers and injects clinical information directly into the generation process. During the training, two pre-trained auxiliary heads, Alzheimer's Disease and Cognitively Normal (AD/CN) classification and brain volume vector regression, provide additional cross entropy and regression losses, ensuring that generated scans remain anatomically plausible and clinically consistent. To sample realistic clinical vectors during inference, we additionally train a diffusion model on clinical vectors, enabling flexible sampling of disease states without the need for manual feature engineering. Experimental results demonstrate high-quality 3D MRI generations. Additionally, adjusting disease labels or specific brain volumes demonstrates a feasible level of conditional control, suggesting that this approach could benefit data augmentation and support clinically relevant neuro-imaging tasks.

1 Introduction

Alzheimer's disease (AD) is widely recognized as the most common cause of dementia worldwide, with recent estimates indicating that 57 million people were living with dementia in 2019, a figure projected to nearly triple by 2050 (20). Due to the complexity of identifying the etiology of AD and the wide range of symptoms, identifying biomarker patterns in the disease's progression would be critical in the long-term treatment of patients (7). Nevertheless, studying the disease's time course and mapping it to certain features requires a large amount of diverse and curated data, which can often be challenging to acquire (19).

Generative AI has gained increasing interest in medical imaging, where data scarcity, privacy concerns, and domain shifts present major obstacles to developing robust deep-learning solutions (26). In particular, 3D MRI datasets can be challenging to acquire and standardize, making it difficult for segmentation or classification models to generalize to unseen clinical contexts. Recent progress in generative modeling-spanning Variational Autoencoders (VAEs) (18), Generative Adversarial Networks (GANs) (5), and Diffusion Models (10), promises data augmentation strategies that can help address these limitations. By leveraging these models, researchers have generated 3D Alzheimer's images to study and predict Alzheimer's disease progression. (22; 16)

[*] R. Marinescu is the corresponding author (ramarine@ucsc.edu).

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

However, purely visual plausibility is insufficient in a medical setting: synthetic volumes must also respect clinical attributes such as diagnostic labels and region-specific volumes. Efforts to incorporate conditional signals have shown that aligning generative processes with meaningful clinical features can yield anatomically consistent outputs. For instance, latent diffusion approaches have recently succeeded in controlling brain MRI generation by specifying factors like patient age or overall brain volume (21). Meanwhile, style-based GAN architectures demonstrate the flexibility to inject global or local style parameters, though many focus on 2D tasks or lack explicit constraints for medical applications.

In this work, we propose a 3D VAE-GAN framework with clinical-style conditioning designed to produce MRI volumes that appear realistic and reflect user-defined brain region volumes and AD or Cognitively Normal (CN) diagnoses. The main contributions of our work can be summarized as follows:

- 3D VAE-GAN with Depthwise-Separable Convolutions: We propose a 3D VAE-GAN framework that integrates a 3D encoder and a clinical-style decoder, both employing depthwise-separable convolutions inspired by the Xception model (3). This approach captures volumetric features more efficiently while reducing the overall parameter count compared to standard 3D convolutions.
- Style-Based Conditioning on Clinical Data: We adopt a style-based approach (inspired by StyleGAN2 (17)) in the decoder, injecting user-defined or automatically sampled clinical attributes (e.g., hippocampus, amygdala, and lateral ventricle volumes, AD/CN label) into each convolutional layer. This design ensures that key morphological details (such as amygdala size or disease status) are accurately reflected in the generated MRI.
- Auxiliary Supervision for Clinical Fidelity: Our pipeline leverages auxiliary supervision from a pre-trained AD/CN classifier and a volume feature regression network. This forces the generator to produce anatomically consistent volumes aligned with diagnostic labels and numerical region-specific measures, enhancing realism and clinical relevance.
- Separate Diffusion Model for Clinical Feature Sampling: By training a diffusion model on the clinical feature space alone, we enable on-demand sampling of realistic volume vectors and labels without hand-crafting each attribute. This approach automatically yields diverse, clinically valid conditions at inference, supporting large-scale synthetic dataset creation.

2 Methodology

An overview of the proposed model is shown in Fig. 1. Our 3D VAE-GAN architecture uses clinical-style conditioning to generate realistic MRI volumes guided by both image data and clinical features (e.g., volumetric measures and AD/CN labels). The system's encoder incorporates image and clinical embeddings to produce a latent distribution, while a conditional prior, relying purely on clinical data, ensures we can sample new MRI volumes at inference without a real image.

2.1 Problem definition and notation

Given a 3D T1-weighted MRI volume $\mathbf{x} \in \mathbb{R}^{H \times W \times D}$ and an associated clinical vector $\mathbf{c}$ (six regional brain volume ratios + an AD/CN label), we seek a conditional generator $p_\theta(\mathbf{x}|\mathbf{c})$ that can draw anatomically realistic MRI scans whose distribution matches that of real data with the same clinical profile.

We introduce a latent code $\mathbf{z} \in \mathbb{R}^k$ and decompose the likelihood as

$$p_\theta(\mathbf{x}|\mathbf{c}) = \int p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c}) \cdot p_\theta(\mathbf{z}|\mathbf{c}) \, d\mathbf{z}. \tag{1}$$

In practice, the integral in Eq. (1) is intractable. Following the standard variational approach, we introduce an encoder $q_\phi(z \mid x, c)$ to approximate the true posterior and optimize the evidence lower bound (ELBO, a standard VAE objective combining reconstruction accuracy and latent regularization). The latent variable z is sampled via the reparameterization trick, $z = \mu_e(x, c) + \sigma_e(x, c)\, \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. During training, we maximize the ELBO by minimizing the reconstruction loss between x and $g_\theta(z, c)$, together with a Kullback–Leibler (KL) divergence term that aligns $q_\phi(z \mid x, c)$ with the

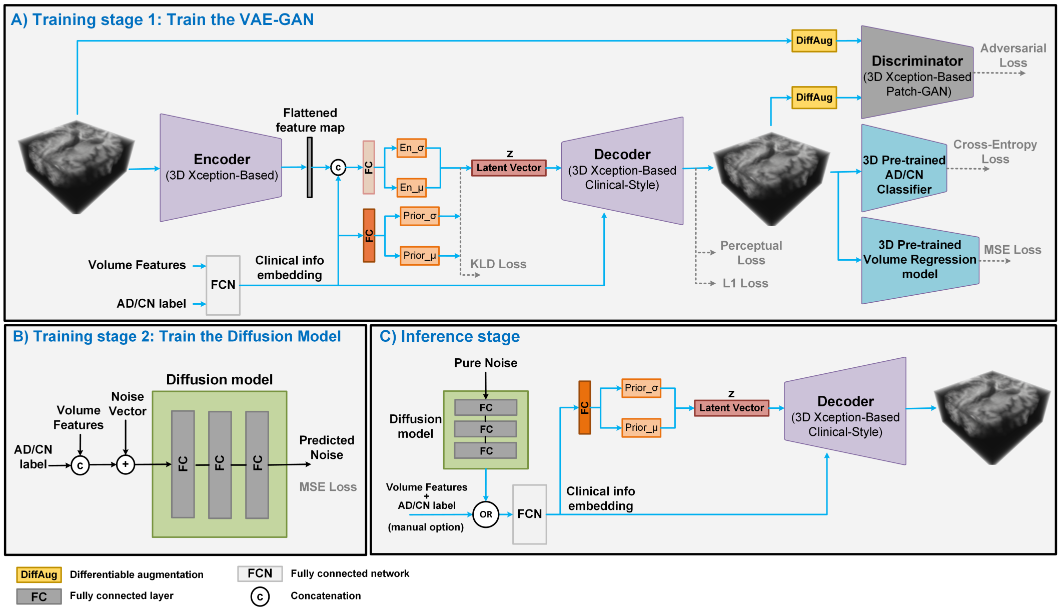

Figure 1: A) VAE-GAN Training: A 3D Xception-based VAE learns latent MRI features. Its style-based decoder is modulated by clinical vectors (volumes + AD/CN). B) Diffusion Model Training: Clinical vectors form a learned distribution for realistic sampling. C) At inference, users can either sample clinical vectors from the diffusion model or provide them manually, then generate synthetic MRI volumes via a conditional prior and the style-based 3D decoder.

learned conditional prior $p_\theta(z \mid c)$. This formulation makes Eq. (1) tractable and allows end-to-end optimization through gradient descent.

- The decoder g_θ implements $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c})$ (right purple block in Fig. 1-A).
- The conditional prior $p_\theta(\mathbf{z}|\mathbf{c})$ is parameterized as a Gaussian $\mathcal{N}\big(\mu_p(c),\, \sigma_p^2(c)\big)$, whose μ_p and σ_p are predicted by a small MLP (FC) that reads embedded c (blocks Prior$_\mu$/Prior$_\sigma$ in Fig. 1-A).
- Separately, we train a diffusion model with parameters ψ on clinical vectors to learn $p_\psi(c)$ (Fig. 1-B). At inference, we either sample $c \sim p_\psi(c)$ or provide c manually (Fig. 1-C), then draw $z \sim p_\theta(z \mid c)$ and decode.

At training time, the encoder e_ϕ maps $(\mathbf{x}, \mathbf{c})$ to the variational posterior $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c})$; we draw $\mathbf{z}$ with the standard re-parameterization trick. For every mini-batch, we minimize the composite loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \tag{2}$$

Table 1: Loss-term definitions

Term	**Definition**
$\mathcal{L}_{\text{rec}}$	L1 voxel-wise error between the ground-truth MRI $\mathbf{x}$ and reconstruction $g_\theta(\mathbf{z}, \mathbf{c})$
$\mathcal{L}_{\text{KL}}$	Kullback-Leibler divergence $\text{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{c}) \,\|\, p_\theta(\mathbf{z} \mid \mathbf{c}))$
$\mathcal{L}_{\text{adv}}$	PatchGAN adversarial loss conditioned on $\mathbf{c}$
$\mathcal{L}_{\text{perc}}$	Perceptual feature loss using pre-trained EfficientNet-B0 layers (25)
$\mathcal{L}_{\text{cls}}$	Auxiliary classification (AD vs CN)
$\mathcal{L}_{\text{reg}}$	ROI-volume regression loss

Each loss plays a different role. $\mathcal{L}_{\text{rec}}$ keeps the reconstruction anatomically aligned with the input and stabilizes training. $\mathcal{L}_{\text{KL}}$ aligns the encoder's latent distribution with the learned clinical prior. $\mathcal{L}_{\text{adv}}$ (PatchGAN) pushes local realism and sharp detail, counteracting the blur that pure reconstruction losses can introduce. $\mathcal{L}_{\mathrm{perc}}$ compares high-level features from pre-trained EfficientNet-B0 (25) rather than raw pixels, encouraging correct global structure and texture. $\mathcal{L}_{\mathrm{cls}}$ enforces that generated images match the AD/CN label in the conditioning vector, while $\mathcal{L}_{\mathrm{reg}}$ keeps the six regional volumes consistent with the requested values.

In all experiments we kept the coefficients fixed at $\lambda_{\text{rec}} = 10$, $\lambda_{\mathrm{KL}} = 0.1$, $\lambda_{\text{adv}} = 1$, $\lambda_{\text{perc}} = 20$, $\lambda_{\text{cls}} = 1$, $\lambda_{\text{reg}} = 1$, a setting that gave us the best balance between visual sharpness (higher λ_{rec}, λ_{perc}) and latent diversity (lower λ_{KL}). These values were selected empirically following evaluation of the results.

2.2 Algorithm Overview

Algorithm 1 VAE–GAN Training Loop (clinical prior)

```
repeat
    Sample $(\mathbf{x}, \mathbf{c})$ from the dataset
    Encode $\to \mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c})$
    Decode $\to \hat{\mathbf{x}} = g_\theta(\mathbf{z}, \mathbf{c})$
    Update encoder and decoder with $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\mathrm{KL}}$, $\mathcal{L}_{\text{perc}}$, $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{reg}}$
    Update discriminator and decoder's adversarial branch with $\mathcal{L}_{\text{adv}}$
until convergence
```

Algorithm 2 Diffusion Training on clinical vectors

```
repeat
    Sample $\mathbf{c}$ from the dataset
    Add Gaussian noise at a random timestep t
    Predict the added noise
    Minimize MSE between the predicted noise and the true noise
until convergence
```

Algorithm 3 Inference (Sampling)

```
Provide or sample $\mathbf{c}$:
    Manual mode: user supplies volume features and/or label
    Stochastic mode: draw $c \sim p_\psi(c)$ using the diffusion model.
Sample    $z \sim p_\theta(\mathbf{z}|\mathbf{c}) = \mathcal{N}\big(\mu_p(c),\, \sigma_p^2(c)\big)$.
Generate $\hat{\mathbf{x}} = g_\theta(\mathbf{z}, \mathbf{c})$, a new MRI that matches $\mathbf{c}$
```

2.3 Datasets

We used T1-weighted MRI scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (1; 15), focusing on participants diagnosed as CN or AD. After acquiring an initial set of scans, we excluded poor-quality data or missing labels, resulting in a final cohort of 2564 MRI volumes. We applied the SynthStrip tool from the FreeSurfer pipeline (12) for skull stripping each MRI scan to remove non-brain tissue, followed by z-score intensity normalization, cropping, and resampling to a consistent 128×128×128 resolution using a third-order B-spline interpolation kernel. In addition to the 3D scans, we used regional volume features for the left/right hippocampus, left/right amygdala, and left/right lateral ventricles, which correlate highly with AD (4). These measurements were normalized by each subject's whole brain volume, creating dimensionless ratios that served as conditioning inputs for both our VAE-GAN and diffusion models. We also used subject-disjoint splits of 80/10/10% (train/val/test). All metrics are reported on the test set.

2.4 Models Architecture

Differentiable Augmentation: To improve generalization and sample diversity, especially on a limited dataset, we adopted a two-stage augmentation pipeline. (i) Conventional 3D transforms. Each training volume was randomly blurred, noised, or gamma adjusted with fixed probabilities, exposing the model to diverse appearances and reducing overfitting. (ii) Differentiable augmentation (28) (DiffAug blocks in Fig. 1-A). Following Zhao et al., we applied differentiable brightness, contrast, and 3-D translations to both real and generated samples during each adversarial update. Because these transforms are parameter–free and differentiable, they preserve gradient flow and stabilize GAN training. Empirically, the combined strategy lowered FID and LPIPS while increasing perceptual diversity.

VAE encoder architecture: The encoder is a 3D Xception-style stack that reduces a $128 \times 128 \times 128$ MRI volume to a compact feature map while keeping compute low. Each stage uses depth-wise separable $3 \times 3 \times 3$ convolutions followed by a $1 \times 1 \times 1$ point-wise convolution (Xception factorization (3)), then down-samples with 3D max-pooling. Channels increase $64 \rightarrow 128 \rightarrow 256 \rightarrow 512$, yielding a feature map at one-eighth spatial resolution (Fig. 2-A). We flatten this map and concatenate (circle ⓒ in (Fig. 1-A)) an MLP embedding of the clinical vector; two linear heads then predict the posterior parameters used in q_{ϕ}(z|x,c). This Xception design replaced heavy full 3D convolutions with channel-wise spatial filtering plus $1 \times 1 \times 1$ mixing, cutting parameters and complexity substantially while preserving receptive-field depth that is critical for 3D volumes.

VAE decoder architecture: The decoder converts the latent vector $\mathbf{z}$ into a full-resolution MRI while injecting clinical context via StyleGAN-inspired modulation (17). The 7–D clinical vector c is mapped by an MLP to a style code w. At each up-sampling stage, we applied a separable depth-wise $3 \times 3 \times 3$ convolution (`sd_c3` in Fig. 2-B) whose weights are modulated by w and then de-modulated similar to StyleGAN2 approach (17). A "style conv" (SC) mixes channels after every up-sample. This sequence, upsample → SC → (`sd_c3` + mod/demod), propagates clinical cues through the hierarchy while keeping the model lightweight. A final $1 \times 1 \times 1$ layer produces the reconstructed 3-D MRI.

Conditional VAE with a learned clinical prior: Following the conditional VAE formulation (24), we replaced the fixed Gaussian prior with a learned, clinical conditioned prior $p_{\theta}(\mathbf{z}|\mathbf{c})$ (23) (Fig. 1-A). Given an MRI volume $\mathbf{x}$ and its clinical vector $\mathbf{c}$, the encoder predicts a mean–variance pair (μ_e, σ_e) that defines the approximate posterior $q_{\phi}(\mathbf{z}|\mathbf{x}, \mathbf{c})$ (Fig. 1-A, blocks En_{μ}/En_{σ}). In parallel, a shallow prior network maps $\mathbf{c}$ to its own parameters (μ_p, σ_p), forming $p_{\theta}(\mathbf{z}|\mathbf{c})$ (blocks Prior_{μ}/Prior_{σ}). During training, we minimized the KL divergence between these two Gaussians, encouraging $\mathbf{z}$ to stay clinically consistent while retaining patient-specific anatomy. At inference (Fig. 1-C), we drew $\mathbf{z} \sim p_{\theta}(\mathbf{z}|\mathbf{c})$, with $\mathbf{c}$ sampled from the diffusion model or provided by the user manually and decoded it into a synthetic MRI.

3D network for clinical-info classification and regression: To enforce clinical consistency, we attached two pre-trained 3D ResNet-18 heads (8), one for AD/CN classification and one for regional volume regression, at the output of the generator (cyan modules in Fig. 1-A). As shown in supplementary Fig. S1-A, both backbones employ depth-wise separable $3 \times 3 \times 3$ convolutions (3), cutting model complexity by $\approx 40\%$ versus standard ResNet-18 while retaining receptive field depth. A global average pooling layer converts the final feature map into a vector that feeds a task-specific fully connected head: (i) two logits with cross-entropy loss for AD/CN discrimination, and (ii) a six-value vector with mean squared error loss for regional volumes. During VAE-GAN training, the weights of both heads were frozen; only their losses propagated to the encoder-decoder, compelling it to synthesize brains that match the conditioning and preserve anatomically plausible volumes. At inference, these auxiliary networks were removed, so generation incurred no extra cost.

Discriminator: As shown in Fig. 1-A and supplementary Fig. S1-B, we used a 3D PatchGAN architecture (14) and replaced its standard convolutions with depth-wise separable $3 \times 3 \times 3$ kernels, reducing parameters and complexity without sacrificing receptive field size. Instead of a single real/fake logit, the network outputs a spatial map of authenticity scores, enabling it to judge local anatomical realism across the entire MRI volume.

Diffusion Model: As shown in Fig. 1-B, we separately trained a diffusion model on the seven-dimensional clinical vector $\mathbf{c}$ (six regional-volume ratios + the AD/CN label). The AD/CN label is encoded as a single binary value (0 = CN, 1 = AD) and concatenated with the six normalized regional-volume ratios to form the 7-D clinical vector c. During sampling, the label value can be fixed to a desired class or sampled jointly with the other features. The model is a three-layer MLP that predicts the added Gaussian noise at each diffusion step; it is optimized with a mean-squared-error loss on the noise term. After training, the diffusion prior was used to sample plausible clinical conditions automatically, a task that would otherwise require labor-intensive feature engineering to maintain realistic co-dependencies among the six volumes and the diagnosis label.

2.5 Training Procedure

All networks were trained in PyTorch from scratch using the Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$). The VAE–GAN generator, encoder, and decoder used a learning rate of 1×10^{-4}, and the discriminator used 1×10^{-3}; the diffusion prior used 5×10^{-4}. Training employed `ReduceLROnPlateau` schedulers (factor $= 0.75$, patience $= 50$–100) and gradient clipping (1.0 for the GAN and diffusion, 5.0 for auxiliary heads). The AD/CN classifier and ROI–volume regression heads were first trained separately on real ADNI images (batch $= 8$, epochs $= 50$–60), then frozen during VAE–GAN training to provide auxiliary losses. The VAE–GAN was trained for 2000 epochs (batch $= 4$) with `ReduceLROnPlateau` on validation FID and early stopping (patience $= 500$). All models were initialized with Xavier uniform weights.

3 Results

3.1 Quantitative evaluation of generated 3D MRI volumes.

We measured the quality of generated 3D MRI volumes using FID (Frechet Inception Distance) (9) as well as SSIM (Structural Similarity Index) (13), and LPIPS (Perceptual Image Patch Similarity) (27). FID was computed on features extracted from a pre-trained 3-D ResNet-18 using 2,000 generated and test-set real volumes. SSIM and LPIPS were evaluated on tri-planar 2-D slices (32 per plane, 96 per volume) and averaged to obtain per-volume scores. The ablation study in Table 2 shows a clear progression: adding the Xception backbone, clinical conditioning, and finally the perceptual loss successively reduces FID and LPIPS while boosting SSIM. Our full model (3D Xception-based VAE-GAN + Clinical conditions + Perceptual) achieved an FID of 30.64, SSIM of 0.89, and LPIPS of 0.23, outperforming all weaker variants. These quantitative gains are qualitatively reflected in Fig. 3, where the generated axial, sagittal, and coronal slices reproduce cortical folding and ventricle anatomy that are visually indistinguishable from the real ADNI scans.

Table 2: Comparison of different methods

Method	#Parameters	FID (↓)	SSIM (↑)	LPIPS (↓)
3D VAE	24.6 M	78.23 ± 1.10	0.72 ± 0.014	0.38 ± 0.016
3D VAE-GAN	35.7 M	53.87 ± 0.95	0.77 ± 0.012	0.31 ± 0.015
3D Xception-based VAE-GAN	7.9 M	47.45 ± 0.82	0.79 ± 0.011	0.28 ± 0.014
3D Xception-based VAE-GAN + Clinical conditions	8.1 M	45.31 ± 0.78	0.80 ± 0.012	0.26 ± 0.013
3D Xception-based VAE-GAN + Clinical conditions + Perceptual loss	13.2 M	**30.64 ± 0.90**	**0.89 ± 0.012**	**0.23 ± 0.014**

3.2 Quantitative comparison with state-of-the-art baselines

To ensure a fair comparison, we re-implemented and trained two strong generative models namely 3D-WGAN-GP (6) and 3D-StyleGAN2 (11) using their publicly available PyTorch codes on the same ADNI train/validation/test split (80/10/10, subject-disjoint), with identical pre-processing and resolution. We measured FID, SSIM, and LPIPS on our ADNI test set. Table 3 shows that our clinically conditioned 3D VAE–GAN achieves the best FID (30.64), highest SSIM (0.89), and lowest LPIPS (0.23) while using only 13.2 M parameters, outperforming 3D–StyleGAN2 by 15 FID points.

3.3 Evaluating Clinical Consistency

To assess how well our synthetically generated MRIs preserve the input clinical attributes, we evaluated both diagnostic labels and ROI volume features. We sampled 1,000 clinical vectors (AD/CN label plus ROI volume ratios) using our diffusion model at inference, which were then decoded into synthetic MRIs with our 3D Clinical-Style decoder. As shown in Table 4, our pre-trained AD/CN classifier predicted the diagnosis of these synthetic images, achieving 80.6% balanced

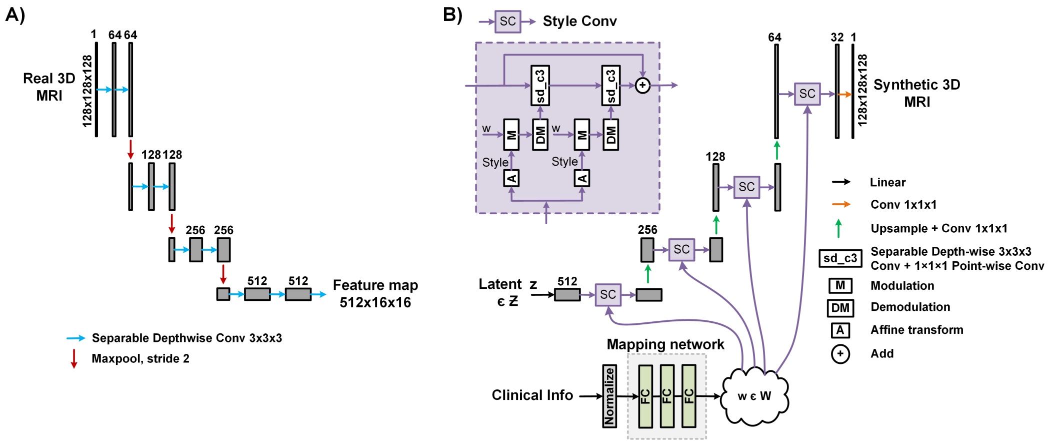

Figure 2: A) 3D Xception encoder. A 3D MRI volume is processed by pairs of depth-wise separable convolutions with max-pooling at each stage, expanding channels and producing a compact feature map. B) Schematic of the 3D clinical-style VAE decoder, highlighting depthwise-separable (sd_c3) convolutions and style-based modulation with clinical embedding.

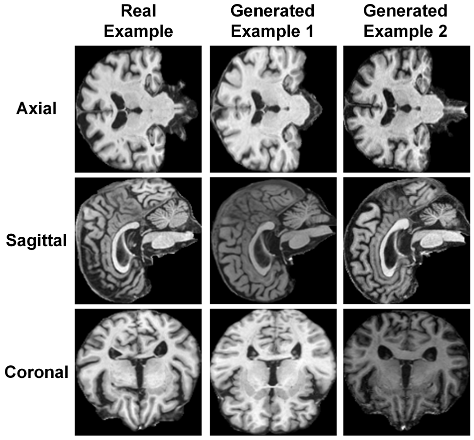

Figure 3: Comparison of real and generated 3D MRI slices (Axial, Coronal, and Sagittal views)

accuracy, close to the 82.1% obtained on real images. Removing the classifier loss ($\mathcal{L}_{\text{cls}}$ in Eq. (2)) during training reduced classification accuracy by 4%. Similar trends were observed with the pre-trained ROI volume regression model: performance dropped when the regression loss ($\mathcal{L}_{\text{reg}}$ in Eq. (2)) was removed.

We further evaluated whether specific subcortical volumes were preserved in the synthetic MRIs. Using SynthSeg tool (2), we segmented the synthetic images to estimate hippocampal, amygdala, and lateral ventricle volumes. Left and right hemisphere measures were summed to obtain total structure volumes, which were then compared to the target values from the input clinical vectors. As shown in

Table 3: Quantitative comparison of 3D brain-MRI generators trained and evaluated on the same ADNI dataset. All baselines were trained with identical pre-processing, resolution (128^3), splits, training budget.

Method	**#Parameters**	**FID (↓)**	**SSIM (↑)**	**LPIPS (↓)**
3D-WGAN-GP (6)	25M	78.2 ± 1.1	0.87 ± 0.01	0.34 ± 0.016
3D-StyleGAN2 (11)	7M	46.0 ± 0.9	0.74 ± 0.011	0.28 ± 0.014
Ours	**13.2M**	**30.64 ± 0.9**	**0.89 ± 0.012**	**0.23 ± 0.014**

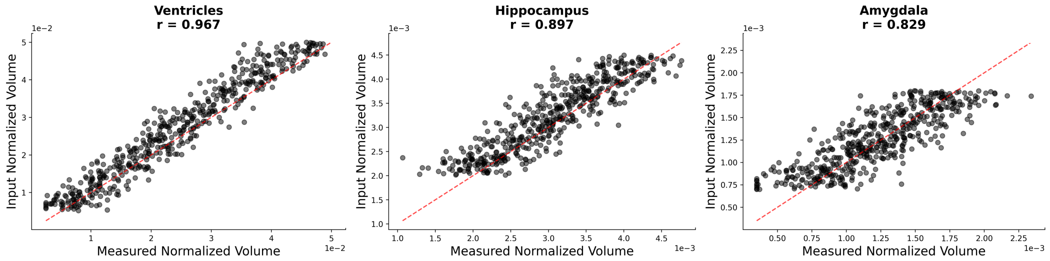

Figure 4: Measured vs. input volumes for the hippocampus, amygdala, and lateral ventricles, extracted from synthetic MRI scans via SynthSeg.

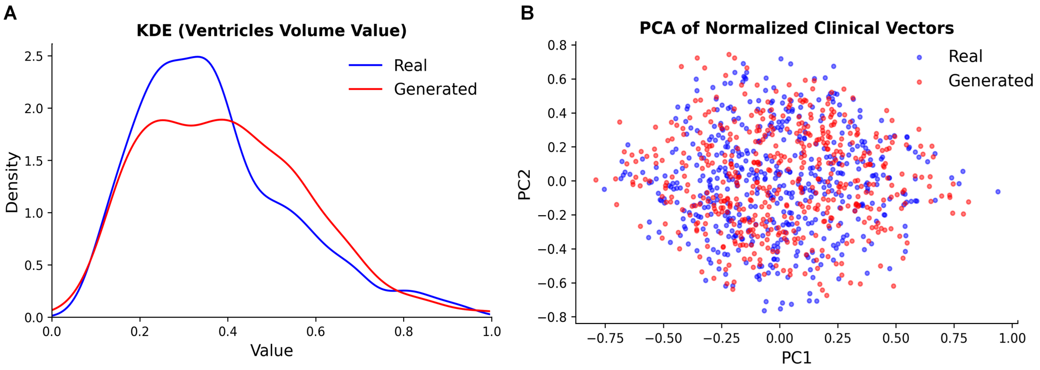

Figure 5: A) KDE for the ventricle volumes, comparing real vs. generated clinical data; and B) PCA of normalized clinical vectors, illustrating the overall distribution overlap between real and generated samples.

Fig. 4, the measured and target volumes exhibited strong correlations, demonstrating that our model effectively preserves clinically relevant structural information in most cases.

Table 4: Comparison of Real and Synthetic 3D Images in Terms of Balanced Accuracy (BAC) for AD/CN Classification and Mean Squared Error (MSE) for Volume Regression of Bilateral Hippocampus, Amygdala, and Ventricle.

	Real	**Gen (with Cls/Reg Loss)**	**Gen (w/o Cls/Reg Loss)**
BAC for AD classification (%)	82.1 ± 1.6	80.6 ± 1.8	76.5 ± 2.1
MSE for volume regression ($\times 10^{-5}$)	3.8 ± 0.3	5.7 ± 0.4	11.1 ± 0.7

3.4 Diffusion Model Evaluation for Clinical Vector Generation

We evaluated the diffusion prior's ability to reproduce the distribution of real clinical variables. Specifically, we compared the marginal distributions of key attributes (e.g., ventricle volumes) between real data and diffusion-sampled clinical vectors. As shown in Fig. 5-A, kernel density estimates (KDE) of real and generated samples overlap closely, indicating that the model captures both the range and shape of the true distribution. We further projected the full 7-D clinical vectors into two dimensions using principal component analysis (PCA). Fig. 5-B shows that real and generated vectors intermix throughout the latent space, suggesting that the diffusion model explores diverse modes rather than collapsing to a narrow subset.

3.5 Computational cost and runtime

To give a practical view of adoption, we reported training times and per sample inference times for each deep learning component in supplementary Table S1. All measurements were taken on a single AWS p3.8xlarge instance (4× V100 16 GB; Intel Xeon E5–2686 v4 @ 2.30 GHz; 244 GB RAM). Supplementary Table S1 summarizes approximate runtimes and hyper parameters. Once trained, deployment is efficient: the VAE GAN synthesizes one MRI in 164 ms (batch=1), making large scale testing feasible. The classifier and regression heads are used only during training (times listed for completeness). The diffusion prior operates on a 7 D vector; with 500 sampling steps it takes about 450 ms to draw one clinical vector, which can be reduced substantially by fewer steps or fast samplers. The end–to–end generation time is therefore 0.61 s per volume when diffusion sampling is used, and 0.16 s when **c** is provided manually.

4 Discussion

We proposed a 3D VAE-GAN with clinical-style conditioning that efficiently captures both visual and anatomical features, enabling fully automated large-scale generation of anatomically and clinically coherent brain scans. On ADNI, the full model reached an FID of 30.64, SSIM of 0.89, and LPIPS of 0.23, outperforming all ablations and similar state-of-the-art baselines. By integrating Xception-based convolutions, adversarial loss, and perceptual loss, we achieved improved image quality and realism. Adding classification and regression constraints during training ensured that synthetic images retained AD/CN classification accuracy and aligned ROI volumes with the input clinical covariates. The diffusion prior model also successfully learned correlations among clinical variables, as shown by KDE and PCA analyses, enabling diverse and realistic sampling without manual feature engineering. The AD/CN classification and ROI regression heads were trained separately on real ADNI data, frozen, and then used only to provide auxiliary supervision during VAE–GAN training. All quantitative evaluations were performed on a held-out test set to avoid data leakage. We acknowledge, however, that evaluating generated images with the same frozen heads may introduce bias, since the generator was partially optimized to satisfy these networks. To address this, we additionally report independent validation using SynthSeg segmentation, which is external to our training process. Also, in practice, we did not observe major failure cases, though extreme or contradictory conditioning values (e.g., unusually large ventricles with an AD label or mismatched volume ratios) occasionally led to anatomically implausible outputs, such as local intensity artifacts or distorted boundaries.

While our approach demonstrates strong realism, it also carries common risks of generative models, such as occasional hallucinated anatomy or spurious correlations learned from biased training data. These should be carefully monitored before downstream clinical use. Also, a key limitation is that processing full 3D volumes remains computationally demanding, even with depth-wise separable convolutions. For example, training on 2,564 MRI scans for 2,000 epochs across four V100 GPUs required 3 days due to memory overhead and floating-point operations needed for 3D data. To address this, we plan to implement mixed precision training for more efficient memory usage and exploit gradient checkpointing to lower memory and time costs. In addition, although our experiments focus on ADNI, the conditioning mechanism is dataset agnostic; future work will test generalization to other available datasets. We will also extend our approach to include Mild Cognitive Impairment (MCI), further broadening the clinical relevance of our synthetic data and supporting more nuanced disease progression studies.

5 Code Availability

The implementation of the proposed framework is publicly available at `https://github.com/NajmehMa/VAE-GAN`.

References

[1] Alzheimer's Disease Neuroimaging Initiative (ADNI): Adni homepage. `http://adni.loni.usc.edu/`, last accessed Nov 2023

[2] Billot, B., Greve, D.N., Puonti, O., Thielscher, A., Van Leemput, K., Fischl, B., Dalca, A.V., Iglesias, J.E., et al.: Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. Medical image analysis **86**, 102789 (2023)

[3] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)

[4] Coupé, P., Manjón, J.V., Lanuza, E., Catheline, G.: Lifespan changes of the human brain in alzheimer's disease. Scientific reports **9**(1), 3998 (2019)

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

[6] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. Advances in neural information processing systems **30** (2017)

[7] Hampel, H., Lista, S., Khachaturian, Z.S.: Development of biomarkers to chart all alzheimer's disease stages: the royal road to cutting the therapeutic gordian knot. Alzheimer's & Dementia **8**(4), 312–336 (2012)

[8] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[9] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

[10] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[11] Hong, S., Marinescu, R., Dalca, A.V., Bonkhoff, A.K., Bretzner, M., Rost, N.S., Golland, P.: 3d-stylegan: A style-based generative adversarial network for generative modeling of three-dimensional medical images. In: MICCAI Workshop on Deep Generative Models. pp. 24–34. Springer (2021)

[12] Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M.: Synthstrip: skull-stripping for any brain image. NeuroImage **260**, 119474 (2022). https://doi.org/https://doi.org/10.1016/j.neuroimage.2022.119474

[13] Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)

[14] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)

[15] Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)

[16] Jung, E., Luna, M., Park, S.H.: Conditional gan with 3d discriminator for mri generation of alzheimer's disease progression. Pattern Recognition **133**, 109061 (2023)

[17] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)

[18] Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)

[19] Kruse, C.S., Goswamy, R., Raval, Y.J., Marawi, S.: Challenges and opportunities of big data in health care: a systematic review. JMIR medical informatics **4**(4), e5359 (2016)

[20] Nichols, E., Steinmetz, J.D., Vollset, S.E., Fukutaki, K., Chalek, J., Abd-Allah, F., Abdoli, A., Abualhasan, A., Abu-Gharbieh, E., Akram, T.T., et al.: Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the global burden of disease study 2019. The Lancet Public Health **7**(2), e105–e125 (2022)

[21] Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. pp. 117–126. Springer (2022)

[22] Ravi, D., Alexander, D.C., Oxtoby, N.P., Initiative, A.D.N.: Degenerative adversarial neuroimage nets: generating images that mimic disease progression. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 164–172. Springer (2019)

[23] Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)

[24] Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems **28** (2015)

[25] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)

[26] Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE transactions on medical imaging **39**(7), 2531–2540 (2020)

[27] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)

[28] Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. Advances in neural information processing systems **33**, 7559–7570 (2020)