AnaDiffusion: Anatomically Compositional Latent Diffusion for Controllable 3D Brain MRI Generation

Huiwen Han[1,2,*,†] **Lulin Liu**[2,3,*,†] **Bangya Liu**[4] **Yuanhao Cai**[5] **Nuo Chen**[2]
Xiaoqing Wang[2] **Ziqian Xie**[6] **Chenyu You**[7] **Shuiwang Ji**[2] **Degui Zhi**[6] **Zhiwen Fan**[2]

[1]Stanford University [2]Texas A&M University [3]University of Minnesota
[4]University of Wisconsin-Madison [5]Johns Hopkins University
[6]Yale University [7]Stony Brook University

[*]Equal contribution [†]Visiting students at Texas A&M University during this work

Abstract

3D brain MRI generation has made significant advances in medical imaging, simulation, and controllable anatomical analysis. However, existing generative models typically synthesize 3D volumes monolithically, often overlooking regional anatomical structures and limiting local controllability. To address these limitations, we introduce **AnaDiffusion**, an anatomically compositional latent diffusion framework that factorizes the generation process into distinct, anatomically meaningful regions, followed by part-to-whole assembly and global refinement. Our approach first trains part diffusion models to capture local structural priors. We then inject an assembled anatomical composite of the parts into the whole-brain latent representation and continue denoising. This mechanism enables the model to resolve global context while preserving the injected anatomy. AnaDiffusion thus produces explicit part assets and a globally coherent volume, enabling controllable part editing without subject-specific dense segmentation maps at inference. On a subject-disjoint ADNI test split, AnaDiffusion attains the lowest mean Fréchet distance (FID) in every evaluated region. In localized editing, it achieves substantially higher target transfer (paired MS-SSIM) than a segmentation-conditioned baseline, with off-target preservation close to the baseline's. Code is available at `https://github.com/phai-lab/AnaDiffusion`, and the project page at `https://anadiffusion.github.io/`.

1 Introduction

High-fidelity 3D brain MRI generation has become increasingly fundamental for medical imaging problems, where large-scale volumetric data are often costly to acquire, difficult to annotate, and limited by privacy and lack of paired anatomical segmentation. Brain MRI synthesis has benefited a wide range of downstream applications, including cross-modality translation [23], super-resolution [22, 49], and disease progression simulation [8, 13, 34, 37]. While recent diffusion models [17, 38] have demonstrated remarkable success in synthesizing globally plausible volumetric data, generating anatomically accurate local structures remains a persistent challenge. Whole-brain samples that appear realistic at a macroscopic scale can exhibit localized anatomical failures–such as blurred tissue interfaces, topological inconsistencies in the cortex, and unnatural asymmetries–as shown in Figure 1. These localized artifacts compromise the downstream utility of the generated data. In practice, they are often exposed by automated segmentation pipelines, which reveal incomplete or structurally inconsistent local anatomy in synthesized volumes [7, 21, 52].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

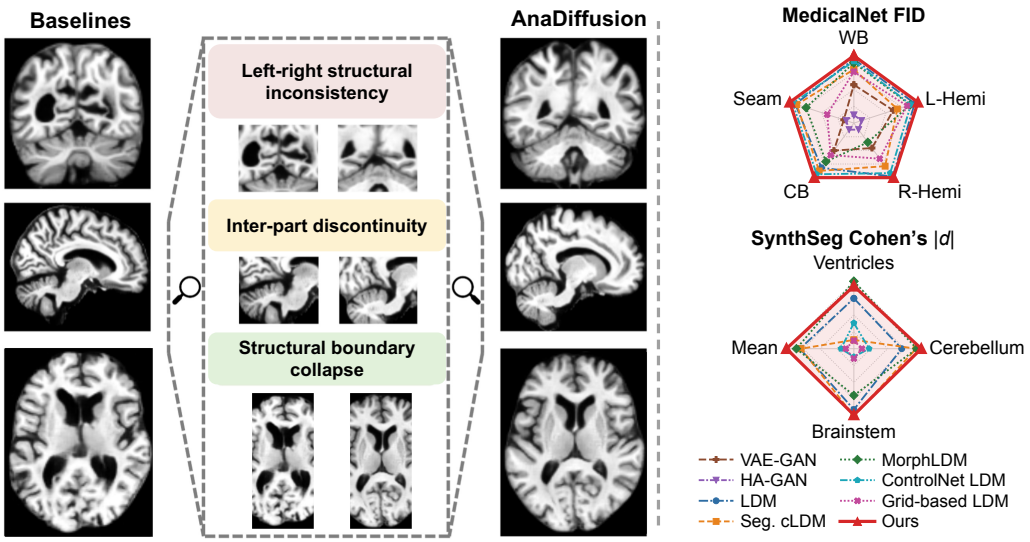

Figure 1: **AnaDiffusion for anatomically compositional 3D brain MRI generation.** Unlike monolithic baselines (left) that may suffer from left-right inconsistency, inter-part discontinuity, and boundary collapse, AnaDiffusion generates part-specific anatomical assets and refines their assembled scaffold into a coherent whole-brain volume. The radar charts (right) summarize Table 1: each axis is min–max normalized on a log scale and lower-is-better metrics (FID, |d|) are inverted, so points farther from the centre are better.

These observations call for 3D MRI generators that preserve fine-grained anatomical structure along with global plausibility. Prior methods add explicit structural constraints. Geometry-driven approaches use deformation fields or surface meshes to encourage plausible shapes [4, 40, 48, 51, 53], and mask-based approaches use segmentation masks as conditioning signals or supervised objectives to enforce a prescribed spatial layout [11, 24, 47]. Mask conditioning gives spatial control but depends on dense input annotations, and the model learns to fill the given boundaries with intensities while the structural dependencies between adjacent regions remain implicit. Both mask-conditioned and unconditioned generators are also *monolithic*: they encode and generate the entire volume through a single, entangled latent representation, although brain regions differ greatly in morphological complexity (e.g., the highly folded cerebral cortex and the densely structured cerebellum) [12]. Monolithic models may therefore average out fine-scale, region-specific details, which motivates a framework that factorizes local structures and integrates them coherently.

Recognizing this gap, we introduce **AnaDiffusion**, an anatomy-informed compositional sampling and refinement framework for 3D brain MRI synthesis. Our contribution lies in how explicit regional generators are coupled to an ongoing whole-brain generative trajectory. To this end, we first train part-specific latent diffusion models, allowing the model to devote generative capacity to anatomically heterogeneous regions. For bilateral structures, we further employ a symmetry-aware hemisphere model with canonicalized weight sharing, which encourages anatomically consistent left-right generation. Building on these local anatomical priors, we then address the central challenge of part-to-whole integration. We construct an anatomical composite from the generated parts and use it to guide a global whole-brain diffusion process, which harmonizes regional details into a coherent volume. To preserve structural consistency, we train the whole-brain denoiser to resume reverse sampling from assembled composite latents reintroduced at intermediate noise levels. We summarize our contributions as follows:

- We introduce **AnaDiffusion**, a compositional latent diffusion framework for controllable 3D brain MRI generation. AnaDiffusion decomposes synthesis into anatomical parts and a whole-brain latent integration process, modeling heterogeneous brain structures more effectively than monolithic whole-volume generation.

- We propose a **part-to-whole latent refinement** that integrates independently generated anatomical parts into a coherent whole-brain volume. This supports controllable part replacement while keeping non-edited regions stable and part-to-whole anatomy consistent.
- Motivated by the approximate structural bilateral symmetry of the human brain, we design a hemisphere generator shared across both hemispheres by employing left-right canonicalization and side-indicator conditioning.
- We demonstrate that, relative to the compared baselines, AnaDiffusion improves the evaluated synthesis metrics on held-out subjects from a subject-disjoint ADNI test split. It achieves stronger regional distributional fidelity and segmentation-based anatomical alignment, while enabling localized part replacement with high target transfer and limited off-target drift, without requiring subject-specific dense segmentation maps at inference time.

2 Related Work

2.1 3D Generative Modeling for Brain MRI

Early work generated brain MRIs slice-wise (2D) or by 2.5D aggregation; more recent work exploits improved compute and architectures for full 3D synthesis, which improves volumetric consistency. GAN-based models such as α-GAN [26, 39], CCE-GAN [54], and HA-GAN [44] produce sharp 3D brain volumes, but their training is prone to instability and mode collapse, which matters when clinically relevant variation is subtle [41]. Diffusion models train more stably. 3D DDPM [9] denoises directly in voxel space at a high computational cost, and latent diffusion models (LDMs) [35] lower this cost by running diffusion in the compact latent space of an autoencoder, which also makes conditioning straightforward [16, 34, 37]. Anatomical plausibility and region-level control remain open problems for these models, and global perceptual metrics such as FID can miss local anatomical errors in generated brain volumes [7, 21, 52].

2.2 Anatomy- and Morphology-Aware 3D Brain Generation

To improve anatomical plausibility, recent 3D brain generators add explicit structural priors. Segmentation maps serve as conditioning signals or auxiliary supervision [2, 4, 10]. For example, Med-DDPM [10] concatenates segmentation masks with the image input to control normal and pathological brain generation, and AG-LDM [47] supervises both the autoencoder and the LDM with tissue- and boundary-aware losses. Morphology-aware methods add geometric constraints, for example template deformation in MorphLDM [48] and conditioned geodesic transformations for aging in IGG [53]. These priors regularize localization, tissue boundaries, or global shape, and each model still generates the brain as a single volume. AnaDiffusion decomposes the brain into anatomical parts with their own generators and assembles them in a whole-brain refinement step, which allows part-level editing without subject-specific dense segmentation maps at inference.

3 Methods

Overview. We introduce AnaDiffusion, a part-to-whole latent diffusion framework for compositional 3D brain MRI generation. AnaDiffusion trains an LDM for each anatomical part and one for the whole brain (Figure 2), then optimizes only the whole-brain denoiser with an assemble-then-refine strategy. Frozen part models generate anatomical parts that are assembled in image space, re-encoded, and injected as a composite latent at t_{inj}. Training optimizes the whole-brain denoiser with standard ϵ-prediction on scaffold-injected latents, together with a seam-band loss that anchors part boundaries to the real scan. At inference, the model produces explicit part assets and a whole-brain volume, enabling part replacement without subject-specific dense segmentation maps.

3.1 Anatomical Part Factorization

Let $x^{\mathrm{wb}} \in \mathbb{R}^{H\times W\times D}$ denote a whole-brain T1 MRI registered to a common template space. We define $P = 3$ anatomical parts: left hemisphere (Hemi_L), right hemisphere (Hemi_R), and cerebellar-brainstem complex (CB). This coarse decomposition follows widely used neuroanatomical organization and is broadly consistent with anatomical and developmental perspectives [14, 28]. Let

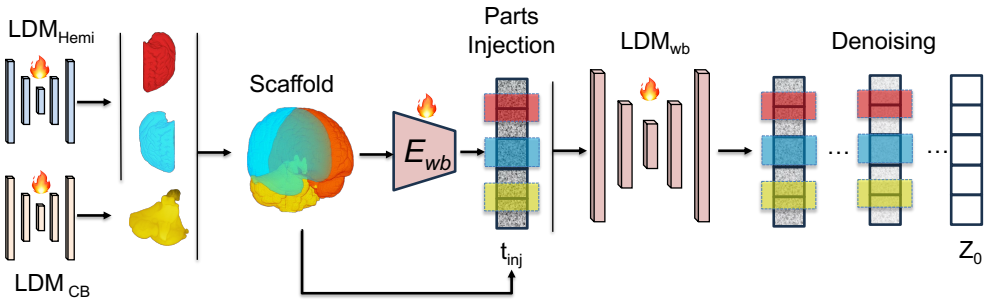

Figure 2: AnaDiffusion performs anatomically compositional 3D brain MRI generation by first synthesizing high-fidelity regional parts and assembling them into a whole-brain scaffold. The scaffold is re-encoded into the whole-brain latent space and injected during denoising, enabling local anatomical detail and globally coherent whole-brain synthesis.

$M^{(p)} \in \{0,1\}^{H\times W\times D}$ be the binary mask for part p, and let $x^{(p)} = \mathcal{C}_p(x^{\mathrm{wb}}) = M^{(p)} \odot x^{\mathrm{wb}}$ denote the corresponding part volume. During training-data preparation, part volumes are extracted from each whole-brain volume by applying part-specific segmentation label maps generated with SynthSeg [3], and then cropped to fixed, part-aligned bounding boxes. These subject-derived labels are used only offline to construct the regional training volumes and are not provided to the diffusion models as subject-specific conditioning inputs. Extra-ventricular cerebrospinal fluid (CSF) is not assigned to any part because it is a fluid compartment whose boundaries are unstable and sensitive to partial-volume effects. We define a deterministic assembly operator $x^{\mathrm{scaf}} = \mathcal{A}(\{x^{(p)}\}_{p=1}^{P})$, which places part assets into their template-aligned locations while omitting CSF, background, and uncertain boundary regions. Details on the anatomical definitions, template masks, crop geometry, and assembly procedure are provided in Appendix A.

3.2 Local Anatomical Priors

For each anatomical support, we train an autoencoder and latent diffusion prior. Let $E^{\mathrm{wb}}, D^{\mathrm{wb}}$ denote the whole-brain encoder and decoder, and $E^{(p)}, D^{(p)}$ denote the corresponding part autoencoder. All autoencoders are trained before diffusion training and then frozen. Given a latent $z_0 = E(x)$, each diffusion denoiser is trained with the standard noise-prediction objective

$$\mathcal{L}_\epsilon = \mathbb{E}_{z_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t,t)\|_2^2\right], \qquad z_t = \alpha_t z_0 + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0,I). \tag{1}$$

Hemisphere canonicalization. The cerebral hemispheres have approximate bilateral correspondence but also contain side-specific variation. We therefore use canonicalization for statistical sharing, not exact symmetry enforcement. Let x^L and x^R denote left- and right-hemisphere crops. We map right-hemisphere crops into a left-oriented canonical frame using a flip operator $\mathcal{F}$, i.e., $\bar{x}^R = \mathcal{F}(x^R)$. A single hemisphere LDM is trained on $\{x^L\} \cup \{\bar{x}^R\}$ and conditioned on a side indicator $y \in \{0,1\}$, where $y = 0$ denotes left and $y = 1$ denotes right. At inference, the model samples in the canonical frame; right-hemisphere samples are flipped back to native orientation using $\mathcal{F}$.

3.3 Part-to-Whole Latent Scaffold

Independently generated parts live in part-specific latent spaces and need not be directly compatible with the whole-brain latent space. We therefore compose in image space and re-encode with the whole-brain encoder. Given generated part assets $\{\tilde{x}^{(p)}\}_{p=1}^{P}$, we assemble $\tilde{x}^{\mathrm{scaf}} = \mathcal{A}(\{\tilde{x}^{(p)}\}_{p=1}^{P})$, then map it into the whole-brain latent space as $z_0^{\mathrm{scaf}} = E^{\mathrm{wb}}(\tilde{x}^{\mathrm{scaf}})$. This avoids direct pasting between incompatible part and whole-brain latent spaces. Each part is pasted into whole-brain coordinates with a fixed map, built from its MNI152 template mask and feathered at the boundary (Appendix A.2). Let m be the downsampled latent-space union of these part-masks in whole-brain coordinates. For a whole-brain latent trajectory z_t^{wb}, we noise the scaffold latent with the trajectory's

own noise at the injection step and replace the content inside the part-masks with the scaffold,

$$z^{\mathrm{inj}}_{t_{\mathrm{inj}}} = z^{\mathrm{wb}}_{t_{\mathrm{inj}}} + \alpha_{t_{\mathrm{inj}}}\, m \odot \big(z_0^{\mathrm{scaf}} - \hat{z}_0\big), \qquad \hat{z}_0 = \frac{z^{\mathrm{wb}}_{t_{\mathrm{inj}}} - \sigma_{t_{\mathrm{inj}}}\, \epsilon_\theta\big(z^{\mathrm{wb}}_{t_{\mathrm{inj}}}, t_{\mathrm{inj}}\big)}{\alpha_{t_{\mathrm{inj}}}}. \tag{2}$$

The whole-brain denoiser then continues reverse sampling from $z^{\mathrm{inj}}_{t_{\mathrm{inj}}}$ to obtain the final latent $\hat{z}_0^{\mathrm{wb}}$, which is decoded as $\hat{x}^{\mathrm{wb}} = D^{\mathrm{wb}}(\hat{z}_0^{\mathrm{wb}})$.

3.4 Assemble-then-Refine Compositional Training

We train the whole-brain denoiser to refine scaffold-injected latents while keeping all part pipelines and the whole-brain autoencoder frozen. For a training scan x^{wb}, let $z_0^{\mathrm{tgt}} = E^{\mathrm{wb}}(x^{\mathrm{wb}})$ be the target whole-brain latent. Frozen part generators produce part assets, which are assembled into $\tilde{x}^{\mathrm{scaf}}$ and re-encoded as z_0^{scaf}. We sample $t = t_{\mathrm{inj}} \sim \mathcal{U}[100, 300]$, noise the target as $z_t = \alpha_t z_0^{\mathrm{tgt}} + \sigma_t \epsilon$, inject the scaffold with Eq. (2) using the stop-gradient estimate $\hat{z}_0$, and minimize

$$\mathcal{L}_{\mathrm{AR}} = \mathcal{L}_\epsilon + \lambda\, \mathbb{E}\Big[w_t \big\| M_{\mathrm{seam}} \odot \big(\epsilon^* - \epsilon_\theta^{\mathrm{wb}}(z_t^{\mathrm{inj}}, t)\big)\big\|_2^2\Big], \qquad \epsilon^* = \frac{z_t^{\mathrm{inj}} - \alpha_t z_0^{\mathrm{tgt}}}{\sigma_t}, \tag{3}$$

where $\mathcal{L}_\epsilon$ is the noise-prediction loss of Eq. (1), M_{seam} is a band around the part boundaries in latent space, $w_t = \min(1, \sigma_t^2/\alpha_t^2)$, and $\lambda = 0.5$. The seam term asks the denoiser to recover the real scan across part boundaries, where the scaffold and the surrounding brain disagree. This objective trains the whole-brain denoiser to continue reverse sampling from a latent state that already contains an assembled anatomical scaffold. The injected scaffold provides explicit regional structure, while the whole-brain denoiser synthesizes omitted regions such as CSF, background, and uncertain boundaries and harmonizes the final global anatomy.

3.5 Part Editing Application

At inference, AnaDiffusion supports two modes. In *unconditional compositional generation*, AnaDiffusion first denoises a whole-brain latent from Gaussian noise to the injection point t_{inj}. The part branches start from the whole-brain sample. We decode the partially denoised whole-brain estimate and use fixed MNI152-derived template masks to localize and crop anatomical regions from it, so each part refinement starts from a globally contextualized coarse structure. Each crop is refined into an explicit part asset by latent-space SDEdit. It is encoded with its part autoencoder, forward-noised to the injection timestep, and denoised by the frozen part LDM. The assets are pasted with the part masks of Section 3.3 and assembled into $\tilde{x}^{\mathrm{scaf}}$ and re-encoded with E^{wb}, re-noised to t_{inj} via Eq. (2), and further denoised to obtain the final coherent volume. No subject-specific dense segmentation map is used at inference.

Part editing is intended for controllable localized synthesis, such as data augmentation for underrepresented regional phenotypes (e.g., unilateral hippocampal atrophy) and counterfactual analyses; clinically specific edits would require regional generators trained with the corresponding condition. In *part editing*, we first generate or select an initial whole-brain sample and its associated part assets. A selected asset, such as the right hemisphere or cerebellar-brainstem complex, may be replaced with an alternative generated part. The modified asset set is then reassembled, re-encoded, injected, and refined by the whole-brain denoiser. This yields a localized edit while preserving the non-target anatomy and repairing seams around the replacement.

During DDIM inference, r_{inj} denotes the number of denoising steps remaining after scaffold injection, which is distinct from the training timestep t_{inj}. Smaller r_{inj} preserves the inserted part more strictly, and larger r_{inj} allows more harmonization (Section 5.3).

4 Data Preprocessing and Anatomical Factorization

Dataset and preprocessing. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`), and Appendix A.4 gives the full ADNI statement. We evaluate on 1,735 T1-weighted brain MRIs from 407 subjects. Subjects are 55-93 years old (mean 76.7), with 47.4% female subjects and diagnostic groups spanning cognitively normal, mild cognitive impairment, and Alzheimer's disease. Each scan

is preprocessed with N4 bias-field correction [46], skull stripping [18], affine registration to MNI152 space [1, 20]*, resampling to $1.5\,\mathrm{mm}$ isotropic resolution, and intensity normalization [42]. To avoid leakage across longitudinal acquisitions, we split the data at the subject level into training, validation, and test sets containing 305, 41, and 61 subjects and 1,288, 156, and 291 scans, respectively. AnaDiffusion was trained on the ADNI training split and evaluated on the subject-disjoint ADNI test split.

Anatomical factorization. We use a coarse, task-driven anatomical factorization into three components: the left hemisphere ($\mathrm{Hemi_L}$), the right hemisphere ($\mathrm{Hemi_R}$), and the cerebellar-brainstem complex (CB). This decomposition follows widely used neuroanatomical organization in neuroimaging and neurobiology and is broadly consistent with anatomical and developmental perspectives [14, 28]. The three parts are interpretable compartments with distinct morphology, the two hemispheres share homologous structure that canonicalization can exploit, and each part is a meaningful unit for localized editing. A recent sparse mixture-of-experts foundation model for multimodal brain MRI also learned experts that specialize to spatially distinct anatomical structures [19], which gives independent support for allocating model capacity by anatomy. SynthSeg labels are used only offline to build the regional training volumes (Section 3.1). Because all volumes are aligned to MNI152 space, fixed template masks give the crop locations and the paste weights at inference, so assembly needs neither subject-specific dense segmentation maps nor additional registration.

5 Experiments

5.1 Experimental Setting

Baselines. We compare AnaDiffusion against seven representative baselines spanning adversarial, diffusion-based, and morphology-based generation: **(i) VAE-GAN** [27], which combines variational autoencoding with adversarial training; **(ii) HA-GAN** [44], a hierarchical adversarial model for high-resolution 3D medical image synthesis; **(iii) a standard whole-brain LDM** [35, 36], trained unconditionally; **(iv) a 3D ControlNet-style LDM** with a matched whole-brain backbone; **(v) a grid-based LDM** that partitions the volume into a $3 \times 3 \times 3$ grid and applies positional encoding to each of the resulting 27 patches, designed to separate the effect of anatomical semantics from generic spatial specialization; **(vi) a segmentation-mask cLDM**, representing prior mask-guided medical diffusion models (e.g., Med-DDPM [10]), which conditions generation on anatomical part masks through multi-channel concatenation; and **(vii) MorphLDM** [48], a template-warping approach that generates subject-specific anatomy by learning deformation fields relative to a canonical template. All baselines use identical subject-level splits and preprocessing, with matched whole-brain backbone configurations where applicable. MorphLDM is the sole preprocessing exception because it requires intensity normalization to $[0, 1]$ rather than $[-1, 1]$.

Implementation. We implement AnaDiffusion in PyTorch using MONAI Generative Models AutoencoderKL and diffusion UNet architectures [5, 36]. Generation resolutions are 128^3 for whole brain, $64 \times 128 \times 128$ for each hemisphere, and $128 \times 96 \times 64$ for the cerebellar–brainstem complex. Autoencoders are trained with AdamW for up to 100 epochs and then frozen for diffusion training. We use a 1,000-step DDPM process with a scaled-linear β schedule ($\beta_{\mathrm{start}} = 0.0015, \beta_{\mathrm{end}} = 0.0195$) and 50-step DDIM sampling at inference. During training, scaffold injection uses $t_{\mathrm{inj}} \sim \mathcal{U}[100, 300]$ on the DDPM schedule. At inference, we inject with $r_{\mathrm{inj}} = 10$ DDIM steps remaining, which corresponds to approximately $t_{\mathrm{inj}} \approx 180$ under the 50-step DDIM subsampling (Appendix A.1).

Metrics. We evaluate generation quality with the Fréchet distance (FID) in the feature space of a 3D ResNet-10 backbone from MedicalNet [6], pre-trained on large-scale medical imaging datasets. FIDs are computed for the whole brain and for template-defined subregions: left hemisphere, right hemisphere, and cerebellar-brainstem complex. Subregion masks are derived from an MNI-space, SynthSeg template and dilated by 2 voxels ($\approx$ 3 mm at 1.5 mm isotropic spacing) to encompass residual anatomical size variation after registration. To quantify interface coherence, we compute the same FID within a fixed seam band around part boundaries. Because FID can miss local anatomical errors, we also assess anatomical plausibility with SynthSeg-derived morphometry by comparing

*We used the Montreal Neurological Institute MNI152 template for image processing purposes; it is available for download at `http://www.bic.mni.mcgill.ca/ServicesAtlases/ICBM152NLin2009`.

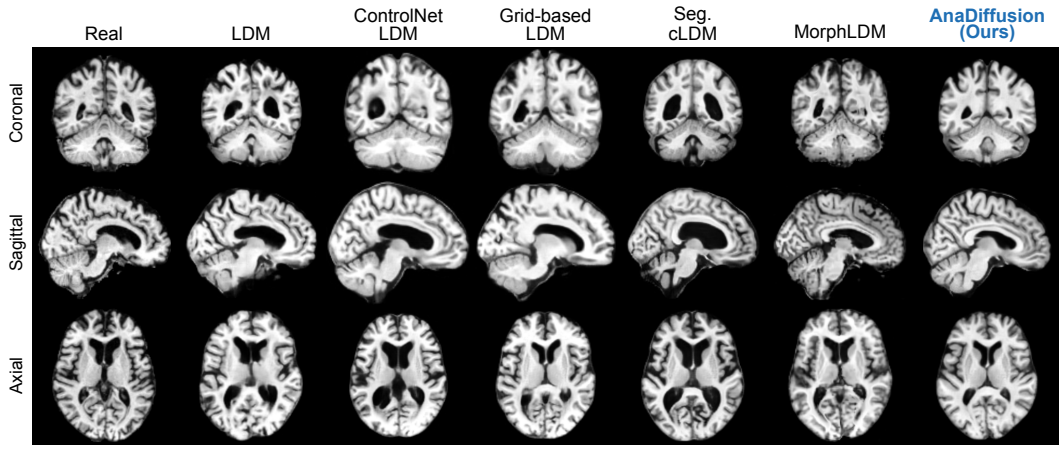

Figure 3: **Qualitative comparison of 3D brain MRI generation.** Columns from left to right show Real, LDM, ControlNet LDM, Grid-based LDM, Seg. cLDM, MorphLDM, and AnaDiffusion (Ours). Monolithic baselines show localized artifacts and lose fine-scale structure, whereas part-to-whole refinement preserves local anatomy and boundary coherence.

real and generated per-structure volume distributions using Cohen's $|d|$ [52]. Each method generates 291 volumes, matching the 291 real test volumes. We report bootstrap means and 95% percentile confidence intervals over 5,000 replicates, each resampling test subjects with replacement (keeping all their longitudinal scans) and, independently, each method's generated volumes.

5.2 Evaluation of Generation Quality

Table 1 shows that AnaDiffusion obtains the lowest mean MedicalNet FID across the whole brain, both hemispheres, the cerebellar–brainstem complex, and seam regions, suggesting improved distributional alignment at both global and part-specific scales. The seam-region gain is consistent with the intended role of whole-brain refinement in harmonizing independently generated parts. For SynthSeg-derived Cohen's $|d|$, AnaDiffusion achieves the lowest cerebellar mismatch, $0.169\,[0.007, 0.482]$, and the second-lowest ventricular and brainstem mismatch, $0.144\,[0.005, 0.416]$ and $0.185\,[0.008, 0.483]$, respectively. These structures cover complementary aspects of anatomy. Ventricles are sensitive to neurodegenerative morphology and lie between parts, while cerebellum and brainstem contain compact structures with fine tissue boundaries and correspond to the dedicated CB part and its boundary with the cerebrum. SynthSeg label maps of AnaDiffusion samples also agree better with real test volumes than those of the LDM baseline in 16 of 17 regions for both Dice and HD95 (Appendix A.3, Table 3), and Figure 5 shows additional samples.

5.3 Ablation Study

We conduct ablation studies to disentangle the contributions of (a) compositional design, (b) latent intervention strength, (c) injection time variation, and (d) assembly space.

Separate Hemisphere Models. We ablate the part generator design by using distinct AEs and LDMs for the left and right hemispheres (instead of a shared hemisphere generator) to evaluate whether part-specific latent spaces improve anatomical fidelity. As shown in Table 2, separate hemisphere models yield lower mean FID for the left hemisphere, $5.59\,[2.10, 10.47]$ versus $6.21\,[3.14, 10.56]$, the cerebellar–brainstem complex, $0.89\,[0.43, 1.54]$ versus $1.06\,[0.64, 1.64]$, and the seam region, $0.19\,[0.10, 0.35]$ versus $0.29\,[0.17, 0.46]$. However, they yield higher mean FID for the right hemisphere, $8.94\,[4.29, 14.76]$ versus $6.67\,[3.67, 10.72]$, and the whole brain, $38.72\,[16.83, 67.13]$ versus $36.16\,[19.16, 59.61]$. The confidence intervals overlap, so neither design is uniformly better. The shared model still gives lower whole-brain FID (36.16 vs. 38.72) and lower mean bilateral FID (6.44 vs. 7.27), which indicates better bilateral balance in the composed brain. In segmentation, CNNs benefit from explicit spatial context in registered brain volumes and can confuse homologous left/right structures without it [31]. Canonicalization with side condition-

Table 1: **Generation and focused morphometric metrics on *ADNI*.** MedicalNet [6] FID is reported as FID $\times 10^4$. SynthSeg [3] metrics report absolute Cohen's $|d|$ for structures most directly tied to part-to-whole integration. WB: whole brain. Hemi: hemisphere. CB: cerebellar-brainstem complex. Seam FID is computed within a fixed seam band. Values are reported as mean [95% percentile bootstrap confidence interval; Section 5.1], and rankings are based on the means. Best results are in **bold**; second-best results are underlined. Lower is better.

Method	MedicalNet FID					SynthSeg Cohen's $\lvert d\rvert$		
	WB	Left Hemi	Right Hemi	CB	Seam	Ventricles	Cerebellum	Brainstem
VAE-GAN [27]	103.54 [75.1, 140.0]	17.21 [13.4, 23.1]	15.35 [11.1, 21.4]	3.73 [2.8, 5.5]	16.79 [12.8, 22.3]	–	–	–
HA-GAN [44]	310.33 [262.9, 360.3]	63.67 [55.7, 72.0]	25.75 [22.0, 30.0]	9.87 [8.3, 11.6]	20.09 [17.1, 23.3]	–	–	–
LDM [35]	40.92 [26.71, 58.86]	7.10 [4.26, 10.85]	6.87 [4.51, 9.57]	1.68 [1.17, 2.29]	0.44 [0.28, 0.63]	0.184 [0.007, 0.523]	0.370 [0.036, 0.747]	0.222 [0.011, 0.540]
Seg. cLDM [10]	59.47 [32.72, 90.83]	14.54 [8.31, 21.92]	9.22 [4.79, 14.59]	1.45 [0.81, 2.30]	0.49 [0.30, 0.73]	0.421 [0.111, 0.725]	0.207 [0.010, 0.509]	**0.178** [0.008, 0.489]
MorphLDM [48]	46.65 [29.85, 70.74]	8.56 [5.90, 12.98]	18.05 [15.54, 21.27]	2.23 [1.88, 2.62]	0.99 [0.72, 1.35]	**0.131** [0.005, 0.374]	0.202 [0.007, 0.510]	0.357 [0.048, 0.662]
ControlNet LDM	44.24 [22.72, 71.79]	8.30 [3.87, 14.15]	7.58 [4.06, 12.19]	1.24 [0.70, 2.00]	0.36 [0.20, 0.57]	0.301 [0.022, 0.665]	1.366 [1.026, 1.761]	1.311 [0.951, 1.667]
Grid-based LDM	64.53 [46.6, 86.0]	9.72 [7.7, 12.8]	11.36 [8.6, 15.4]	3.02 [2.2, 4.5]	4.83 [3.3, 7.3]	0.427 [0.082, 0.757]	1.821 [1.449, 2.265]	1.235 [0.888, 1.591]
Ours	**36.16** [19.16, 59.61]	**6.21** [3.14, 10.56]	**6.67** [3.67, 10.72]	**1.06** [0.64, 1.64]	**0.29** [0.17, 0.46]	0.144 [0.005, 0.416]	**0.169** [0.007, 0.482]	0.185 [0.008, 0.483]

ing may likewise share a prior across homologous anatomy, with the side indicator supplying lateral context without enforcing mirror symmetry. We hypothesize a sample-efficiency and regularization trade-off. The shared model pools observations from both sides, while separate models have more side-specific capacity but fewer training examples each and may show more optimization, sampling, or finite-sample FID variability. Direction-dependent effects of representational ordering, reported for sequence models [15, 25], motivate further study but do not directly explain our convolutional U-Net results. We present these explanations as hypotheses, since attributing the asymmetry to biological laterality would require targeted experiments.

Latent Injection. We evaluate the contribution of latent injection by comparing full scaffold injection ($\alpha = 1$) with no injection ($\alpha = 0$). As shown in Table 2, disabling injection increases the mean FID across every evaluated region, including the whole brain, from $36.16\,[19.16, 59.61]$ to $52.27\,[30.69, 79.63]$, and the seam region, from $0.29\,[0.17, 0.46]$ to $0.55\,[0.35, 0.81]$. This suggests that injecting the composed part latents provides useful local anatomical priors for regional fidelity and whole-brain composition.

Injection Time Variation. We vary r_{inj} at inference to quantify the trade-off between local part anchoring and global harmonization. Here, r_{inj} denotes the number of DDIM steps remaining after part-latent injection. Since training samples injection timesteps from $t_{\text{inj}} \in [100, 300]$ on the 1000-step DDPM schedule, we evaluate $r_{\text{inj}} \in \{7, 10, 15\}$, corresponding approximately to $t_{\text{inj}} \in \{120, 180, 280\}$. As shown in Table 2, $r_{\text{inj}} = 10$ yields the lowest mean FID across all five evaluated regions, although the confidence intervals overlap. This point-estimate pattern is consistent with a balance between retaining enough refinement for seam harmonization and preventing subsequent denoising from weakening the injected part structure.

Assembly Space. Because the regional and whole-brain autoencoders are trained independently, their latent spaces are not necessarily aligned, even where they have the same dimensionality. AnaDiffusion therefore assembles the decoded part assets in image space and re-encodes the scaffold with the whole-brain encoder. Assembling the part latents without re-encoding raises the mean FID in every region (Table 2), from 36.16 to 149.05 for the whole brain and from 1.06 to 18.23 for the cerebellar–brainstem complex, and produces intensity shifts and localized structural failures.

Table 2: **Ablation study on *ADNI*.** We evaluate the effects of compositional design, latent injection strength, injection timing, and assembly space (image-space assembly with re-encoding, as in AnaDiffusion, versus direct assembly of the part latents) on generation quality. Values are MedicalNet FID $\times 10^4$ for the whole brain (WB), left/right hemispheres, cerebellar-brainstem complex (CB), and seam region. Best results are **bold** and second-best are underlined (lower is better).

Method	WB FID	Left Hemi FID	Right Hemi FID	CB FID	Seam FID
Compositional design ($r_{\text{inj}} = 10$)					
separate hemisphere models	38.72 [16.83, 67.13]	**5.59 [2.10, 10.47]**	8.94 [4.29, 14.76]	**0.89 [0.43, 1.54]**	**0.19 [0.10, 0.35]**
Latent injection strength ($r_{\text{inj}} = 10$)					
w/o latent injection ($\alpha = 0$)	52.27 [30.69, 79.63]	9.83 [5.40, 15.53]	8.52 [5.23, 12.55]	1.84 [1.12, 2.75]	0.55 [0.35, 0.81]
full latent injection ($\alpha = 1$) (Ours)	**36.16 [19.16, 59.61]**	6.21 [3.14, 10.56]	**6.67 [3.67, 10.72]**	1.06 [0.64, 1.64]	0.29 [0.17, 0.46]
Injection timing					
$r_{\text{inj}} = 7$	53.27 [34.21, 76.75]	9.31 [5.88, 13.73]	9.04 [5.88, 12.76]	1.75 [1.22, 2.41]	0.69 [0.51, 0.91]
$r_{\text{inj}} = 15$	51.66 [28.31, 80.82]	8.96 [4.73, 14.63]	9.52 [5.39, 14.82]	1.43 [0.90, 2.10]	0.31 [0.16, 0.51]
Assembly space ($r_{\text{inj}} = 10$)					
w/o re-encoding (latent assembly)	149.05 [104.63, 199.33]	26.77 [16.99, 37.68]	23.93 [15.95, 32.99]	18.23 [16.20, 20.36]	0.63 [0.45, 0.86]

5.4 Controllable Part Editing and Compositional Refinement

We evaluate localized editing using paired multi-scale structural similarity (MS-SSIM) [50]. For each replacement region, we compare AnaDiffusion with the Seg. cLDM baseline using the same donor-recipient editing pairs. A recipient whole brain is first sampled from the standard LDM, and an independent donor part is sampled from the corresponding part model. We replace the recipient target part with the donor part, keep all other parts fixed, and then assemble, re-encode, inject, and refine the edited asset with the AnaDiffusion whole-brain denoiser. For the Segm. cLDM baseline, we use an SDEdit-style editing protocol [30] that updates the whole-brain conditioning mask with the selected donor part. We generate $N = 100$ edited scans from 48 subjects per region and report mean MS-SSIM with 95% confidence intervals from 5,000 subject-cluster bootstrap replicates.

Figure 4a summarizes the quantitative editing results. Target transfer measures whether the inserted donor part is preserved in the final edited brain, while off-target preservation measures whether the non-edited anatomy remains unchanged. AnaDiffusion achieves substantially higher target transfer than Segm. cLDM for the left hemisphere (0.9427 vs. 0.7952), right hemisphere (0.9484 vs. 0.8343), and cerebellar–brainstem complex (0.9618 vs. 0.9151). Off-target preservation remains consistently high for AnaDiffusion (0.9363–0.9496), although it is slightly lower than for Segm. cLDM (0.9447–0.9512). AnaDiffusion also yields larger transfer gains and positive locality contrasts across all three regions, indicating that the edits move the target anatomy toward the donor while concentrating changes within the intended region. The qualitative example in Figure 4b further shows that the substituted part can be inserted into the recipient brain without disrupting the surrounding anatomy.

6 Conclusion

In this work, we presented AnaDiffusion, a part-to-whole latent diffusion framework for controllable 3D brain MRI generation. The central motivation behind AnaDiffusion is that brain MRI synthesis should not be treated purely as monolithic volume generation: anatomical regions have distinct local distributions, interact through structured interfaces, and often require region-specific control. By explicitly generating anatomical part assets and refining their assembled scaffold with a whole-brain denoiser, AnaDiffusion introduces anatomical compositionality into latent diffusion generation. The output is a coherent whole-brain sample with reusable, editable anatomical components. AnaDiffusion attains the lowest mean FID in every evaluated region and clearly higher target transfer in localized editing than a segmentation-conditioned baseline, without subject-specific dense segmentation masks at inference; ablations suggest that latent injection, the injection schedule, and image-space assembly contribute to these results.

Limitations. Despite these advantages, several limitations remain. First, the multi-stage design introduces additional training and inference complexity compared with a monolithic LDM. Second, the current fixed, anatomy-driven factorization improves local part and seam behavior, but does not fully solve broader tissue-level calibration. Third, the chosen decomposition focuses on predefined

(a) Quantitative editing locality

Region	Method	Target transfer ↑	Off-target preserv. ↑	Transfer gain ↑	Locality contrast ↑
Left hemisphere	Ours	**0.9427 [0.9411, 0.9441]**	0.9371 [0.9359, 0.9383]	**0.1627 [0.1565, 0.1688]**	**0.1207 [0.1151, 0.1261]**
	Segm. cLDM	0.7952 [0.7887, 0.8018]	**0.9487 [0.9478, 0.9496]**	0.0152 [0.0138, 0.0168]	−0.0078 [−0.0090, −0.0065]
Right hemisphere	Ours	**0.9484 [0.9471, 0.9495]**	0.9363 [0.9350, 0.9376]	**0.1372 [0.1313, 0.1428]**	**0.1118 [0.1064, 0.1172]**
	Segm. cLDM	0.8343 [0.8300, 0.8384]	**0.9447 [0.9434, 0.9458]**	0.0231 [0.0207, 0.0254]	0.0056 [0.0036, 0.0076]
CB complex	Ours	**0.9618 [0.9609, 0.9628]**	0.9496 [0.9484, 0.9508]	**0.0752 [0.0716, 0.0788]**	**0.0313 [0.0278, 0.0348]**
	Segm. cLDM	0.9151 [0.9122, 0.9183]	**0.9512 [0.9500, 0.9523]**	0.0285 [0.0269, 0.0302]	−0.0207 [−0.0215, −0.0199]

(b) Example localized part replacement

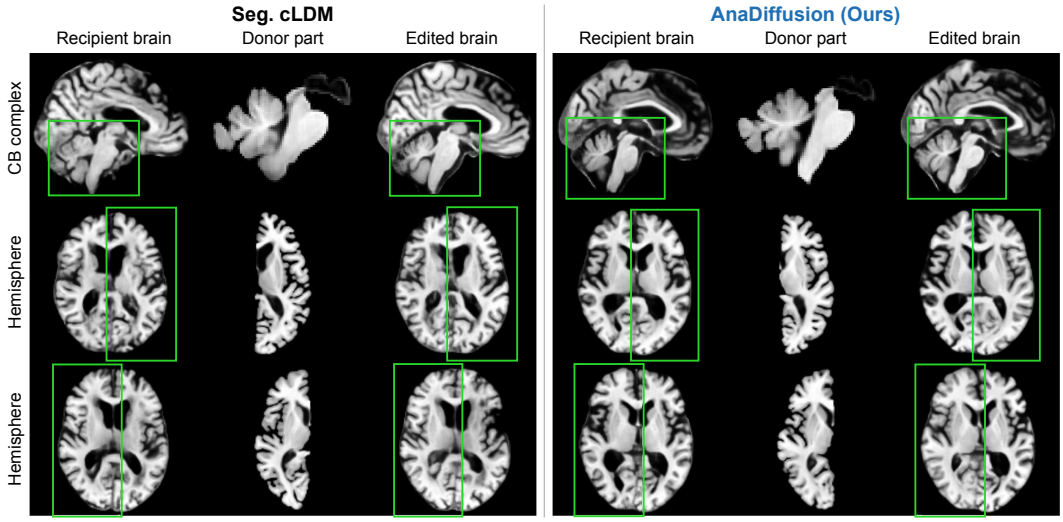

Figure 4: **Localized part editing.** We evaluate whether part replacement edits are both effective and spatially localized. **(a)** AnaDiffusion and the SDEdit-style Segm. cLDM baseline are evaluated on the same $N = 100$ donor-recipient edits from 48 subjects per region. Values are mean paired MS-SSIM with 95% confidence intervals from 5,000 subject-cluster bootstrap replicates. Target transfer measures similarity between the edited target region and the inserted donor part, while off-target preservation measures similarity outside the edited region before and after editing. Contrast metrics further summarize the transfer gain and locality contrast. **(b)** A donor part asset is substituted into a generated recipient brain and refined into an edited whole-brain sample.

anatomical regions and may not capture other biologically meaningful organizations, such as tissue classes, functional networks, or multi-scale anatomical hierarchies. Fourth, AnaDiffusion assumes approximate correspondence with MNI152 and has not been validated for severe mass effect or displaced anatomical boundaries. Such pathology may impair registration and invalidate fixed regional placements. Supporting these cases would require lesion-aware or subject-adaptive localization.

Future work. Future work will explore learned, adaptive decompositions beyond the three coarse parts, such as finer subcortical or lobar units, and will compare them directly with anatomy-free factorizations. Assembly could move into latent space through a shared autoencoder or a learned cross-latent mapping, which would remove the decode and re-encode step and reduce the cost of the multi-stage pipeline. Region-dependent refinement schedules may let the whole-brain denoiser spend more steps on parts and seams with complex morphology. The modular design also offers a path toward condition-specific synthesis. A selected part generator could be trained or fine-tuned with a clinical or morphological condition, such as age, diagnosis, or localized atrophy, and its output inserted and reconciled through whole-brain refinement. Finally, we will train segmentation models on generated volumes and evaluate the approach on other cohorts, scanners, pathologies, resolutions, and imaging contrasts, including younger populations.

Acknowledgments and Disclosure of Funding

This work was supported in part by GPU hardware and cloud credits provided to the Texas A&M University PHAI Lab through the NVIDIA Academic Grant Program. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing, including VISION GPU Cluster. Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Numbers U01AG070112 and R01AG081398. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors declare no competing interests.

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative* (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19 AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

References

[1] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1):26–41, 2008.

[2] Moinak Bhattacharya, Saumya Gupta, Annie Singh, Chao Chen, Gagandeep Singh, and Prateek Prasanna. BrainMRDiff: A diffusion model for anatomically consistent brain MRI synthesis. *arXiv preprint arXiv:2504.04532*, 2025.

[3] Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.

[4] Fabian Bongratz, Yitong Li, Sama Elbaroudy, and Christian Wachinger. Cortex-grounded diffusion models for brain image generation. *arXiv preprint arXiv:2601.19498*, 2026.

[5] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

[6] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer learning for 3D medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[7] Yash Deo, Yan Jia, Toni Lassila, William AP Smith, Tom Lawton, Siyuan Kang, Alejandro F Frangi, and Ibrahim Habli. Metrics that matter: Evaluating image quality metrics for medical image generation. *arXiv preprint arXiv:2505.07175*, 2025.

[8] Nikhil J Dhinagar, Sophia I Thomopoulos, Emily Laltoo, and Paul M Thompson. Counterfactual MRI generation with denoising diffusion models for interpretable Alzheimer's disease effect detection. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–6. IEEE, 2024.

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`

[9] Zolnamar Dorjsembe, Sodtavilan Odonchimed, and Furen Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical imaging with deep learning*, 2022.

[10] 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, 2024.

[11] Virginia Fernandez, Walter Hugo Lopez Pinaya, Pedro Borges, Mark S Graham, Petru-Daniel Tudosiu, Tom Vercauteren, and M Jorge Cardoso. Generating multi-pathological and multi-modal images and labels for brain MRI. *Medical Image Analysis*, 97:103278, 2024.

[12] Bruce Fischl, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355, 2002.

[13] Jingru Fu, Yuqi Zheng, Neel Dey, Daniel Ferreira, and Rodrigo Moreno. Synthesizing individualized aging brains in health and disease with generative models and parallel transport. *Medical Image Analysis*, 105:103669, 2025.

[14] Laura Gui, Radoslaw Lisowski, Tamara Faundez, Petra S Hüppi, François Lazeyras, and Michel Kocher. Morphology-driven automatic segmentation of MR images of the neonatal brain. *Medical image analysis*, 16(8):1565–1579, 2012.

[15] Osama Hardan, Omar Elshenhabi, Tamer Khattab, and Mohamed Mabrok. Flatten wisely: How patch order shapes Mamba-powered vision for MRI segmentation. *arXiv preprint arXiv:2507.13384*, 2025. doi: 10.48550/arXiv.2507.13384.

[16] Miguel Herencia García del Castillo, Ricardo Moya Garcia, Manuel Jesús Cerezo Mazón, Ekaitz Arriola Garcia, and Pablo Menéndez Fernández-Miranda. Diffusion models for conditional MRI generation. *arXiv preprint arXiv:2502.18620*, 2025. URL `https://arxiv.org/abs/2502.18620`.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[18] Andrew Hoopes, Jocelyn S Mora, Adrian V Dalca, Bruce Fischl, and Malte Hoffmann. SynthStrip: skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022.

[19] Haoxu Huang, Long Chen, Jingyun Chen, Jinu Hyun, James Ryan Loftus, Kara Melmed, Daniel Orringer, Jennifer Frontera, Seena Dehkharghani, Arjun Masurkar, and Narges Razavian. Learning sparse latent predictive foundation model for multimodal neuroimaging. *arXiv preprint arXiv:2606.14957*, 2026.

[20] Juan Eugenio Iglesias. A ready-to-use machine learning tool for symmetric multi-modality registration of brain MRI. *Scientific Reports*, 13(1):6657, 2023.

[21] Bahram Jafrasteh, Wei Peng, Cheng Wan, Yimin Luo, Ehsan Adeli, and Qingyu Zhao. WASABI: A metric for evaluating morphometric plausibility of synthetic brain MRIs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 684–694. Springer, 2025.

[22] Mohammad Khateri, Serge Vasylechko, Morteza Ghahremani, Liam Timms, Deniz Kocanaogullari, Simon K Warfield, Camilo Jaimes, Davood Karimi, Alejandra Sierra, Jussi Tohka, et al. MRI super-resolution with deep learning: A comprehensive survey. *arXiv preprint arXiv:2511.16854*, 2025.

[23] Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3D medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter conference on applications of computer Vision*, pages 7604–7613, 2024.

[24] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2024.

[25] Declan Kutscher, David M. Chan, Yutong Bai, Trevor Darrell, and Ritwik Gupta. REOrdering Patches Improves Vision Models. In *Advances in Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=g56WiaXKGF`.

[26] Gihyun Kwon, Chihye Han, and Dae-shik Kim. Generation of 3D brain MRI using auto-encoding generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 118–126. Springer, 2019.

[27] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, 2016.

[28] Lichen Liang, Kelly Rehm, Roger P Woods, and David A Rottenberg. Automatic segmentation of left and right cerebral hemispheres from MRI brain volumes using the graph cuts algorithm. *NeuroImage*, 34 (3):1160–1170, 2007.

[29] John C Mazziotta, Arthur W Toga, Alan Evans, Peter Fox, and Jack Lancaster. A probabilistic atlas of the human brain: Theory and rationale for its development: The international consortium for brain mapping (icbm). *Neuroimage*, 2(2):89–101, 1995.

[30] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=aBsCjcPu_tE`.

[31] Philip Novosad, Vladimir Fonov, D Louis Collins, and Alzheimer's Disease Neuroimaging Initiative. Accurate and robust segmentation of neuroanatomy in T1-weighted MRI by combining spatial priors with deep convolutional neural networks. *Human brain mapping*, 41(2):309–327, 2020.

[32] NVIDIA Corporation. CUDA C++ programming guide. `https://docs.nvidia.com/cuda/archive/12.9.1/cuda-c-programming-guide/index.html`, 2025. Version 12.9.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[34] Wei Peng, Ehsan Adeli, Tomas Bosschieter, Sang Hyun Park, Qingyu Zhao, and Kilian M Pohl. Generating realistic brain MRIs via a conditional diffusion probabilistic model. In *International conference on medical image computing and computer-assisted intervention*, pages 14–24. Springer, 2023.

[35] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI workshop on deep generative models*, pages 117–126. Springer, 2022.

[36] Walter HL Pinaya, Mark S Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F Da Costa, Ashay Patel, et al. Generative AI for medical imaging: Extending the MONAI framework. *arXiv preprint arXiv:2307.15208*, 2023.

[37] Lemuel Puglisi, Daniel C Alexander, and Daniele Ravì. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 173–183. Springer, 2024.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[39] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017.

[40] Filip Rusak, Rodrigo Santa Cruz, Léo Lebrat, Ondrej Hlinka, Jurgen Fripp, Elliot Smith, Clinton Fookes, Andrew P Bradley, Pierrick Bourgeat, Alzheimer's Disease Neuroimaging Initiative, et al. Quantifiable brain atrophy synthesis for benchmarking of cortical thickness estimation methods. *Medical Image Analysis*, 82:102576, 2022.

[41] Muhammad Muneeb Saad, Ruairi O'Reilly, and Mubashir Husain Rehmani. A survey on training challenges in generative adversarial networks for biomedical image analysis. *Artificial Intelligence Review*, 57(2):19, 2024.

[42] Russell T Shinohara, Elizabeth M Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J Mateen, Peter A Calabresi, Samson Jarso, Dzung L Pham, Daniel S Reich, Ciprian M Crainiceanu, et al. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19, 2014.

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[44] Li Sun, Junxiang Chen, Yanwu Xu, Mingming Gong, Ke Yu, and Kayhan Batmanghelich. Hierarchical amortized GAN for 3D high resolution medical image synthesis. *IEEE journal of biomedical and health informatics*, 26(8):3966–3975, 2022.

[45] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, 2015. doi: 10.1186/s12880-015-0068-x.

[46] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: improved N3 bias correction. *IEEE transactions on medical imaging*, 29(6): 1310–1320, 2010.

[47] Cheng Wan, Bahram Jafrasteh, Ehsan Adeli, Miaomiao Zhang, and Qingyu Zhao. Anatomically guided latent diffusion for brain MRI progression modeling. *arXiv preprint arXiv:2601.14584*, 2026.

[48] Alan Q Wang, Fangrui Huang, Bailey Trang, Wei Peng, Mohammad Abbasi, Kilian Pohl, Mert R Sabuncu, and Ehsan Adeli. Generating novel brain morphology by deforming learned templates. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 207–217. Springer, 2025.

[49] Jueqi Wang, Jacob Levman, Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, M Jorge Cardoso, and Razvan Marinescu. InverseSR: 3D brain MRI super-resolution using a latent diffusion model. In *International conference on medical image computing and computer-assisted intervention*, pages 438–447. Springer, 2023.

[50] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003.

[51] Matthias Wilms, Jordan J Bannister, Pauline Mouches, M Ethan MacDonald, Deepthi Rajashekar, Sönke Langner, and Nils D Forkert. Invertible modeling of bidirectional relationships in neuroimaging with normalizing flows: application to brain aging. *IEEE Transactions on Medical Imaging*, 41(9):2331–2347, 2022.

[52] Jiaqi Wu, Wei Peng, Binxu Li, Yu Zhang, and Kilian M Pohl. Evaluating the quality of brain MRI generators. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15010, pages 297–307. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72117-5_28.

[53] Nian Wu, Nivetha Jayakumar, Jiarui Xing, and Miaomiao Zhang. IGG: Image generation informed by geodesic dynamics in deformation spaces. In *International Conference on Information Processing in Medical Imaging*, pages 232–246. Springer, 2025.

[54] Shibo Xing, Harsh Sinha, and Seong Jae Hwang. Cycle consistent embedding of 3D brains with auto-encoding generative adversarial networks. In *Medical Imaging with Deep Learning*, 2021.

A Technical Appendices and Supplementary Material

A.1 Training and Inference

We adopt the AutoencoderKL and diffusion UNet architectures from the MONAI Generative Models library [36], built on MONAI [5]. All models are implemented in PyTorch [33] with CUDA [32] and trained with the AdamW optimizer. The final generation resolution is $128 \times 128 \times 128$ for the whole brain, $64 \times 128 \times 128$ for each hemisphere, and $128 \times 96 \times 64$ for the cerebellar-brainstem complex. For autoencoder training, we use a learning rate of 1×10^{-4} for the whole-brain model and 5×10^{-5} for part models. We warm-start training for 20 epochs before enabling adversarial supervision using a patch-based discriminator, with the discriminator learning rate set to half of the corresponding autoencoder learning rate. Autoencoders are trained for up to 100 epochs, and we select the checkpoint with the lowest exponential moving average (EMA) of validation $L1$ reconstruction loss. For diffusion training, all autoencoders are frozen. All diffusion UNet models are trained for 140,000 optimizer steps. We use a 1,000-step DDPM [17] process with a scaled-linear β schedule ($\beta_{\text{start}} = 0.0015$, $\beta_{\text{end}} = 0.0195$) and DDIM [43] sampling with 50 steps at inference. For injecting part latents, we sample $t_{\text{inj}} \in [100, 300]$ during training to balance (i) having sufficient global structure in the noisy latent for contextual cues and (ii) retaining enough remaining denoising steps for the injected constraint to propagate. Injecting substantially earlier can weaken part identity, while injecting too late provides insufficient propagation. At inference, we inject with $r_{\text{inj}} = 10$ DDIM steps

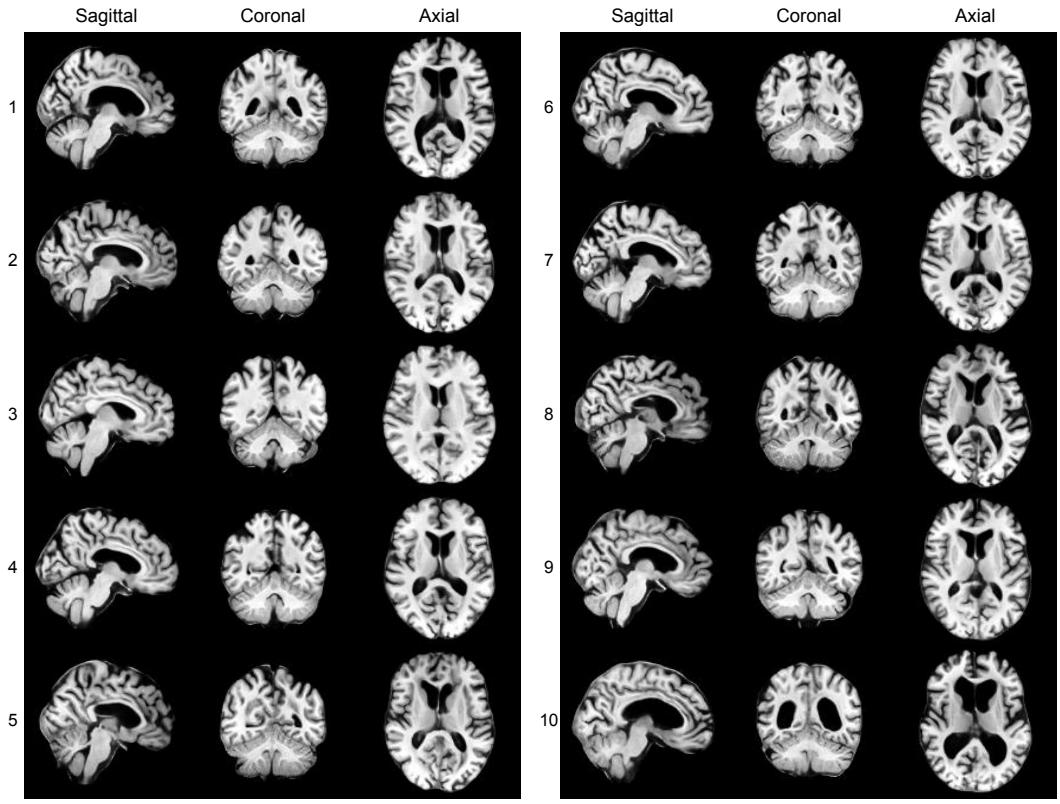

Figure 5: **Additional AnaDiffusion samples.** Ten representative generated samples, each shown as sagittal, coronal and axial slices (slice indices 59, 40 and 65).

remaining, which corresponds to approximately $t_{\text{inj}} \approx 180$ under the 50-step DDIM subsampling. This matches the training noise regime while retaining enough reverse steps for part propagation and seam harmonization. All models were trained on H200 GPUs. Part refinement uses latent-space SDEdit with 100 DDIM steps. During training, the part models start from a timestep five times the injection timestep, capped at $T = 1000$ (500–999 for injection timesteps in $[100, 300]$); at inference, they start from the injection timestep. The stronger training perturbation exposes the whole-brain denoiser to large part–scaffold misalignments and keeps the parts from reproducing the whole-brain model's own output. The seam loss uses a band of two latent voxels on each side of the part boundaries, a weight of 0.5, and the noise-level weighting of Section 3.4. FID features come from the MedicalNet ResNet-10 weights released on PyTorch Hub (`Warvito/MedicalNet-models`), with intensities z-scored within each region; for evaluation, the seam band has a half-width of 2 voxels.

A.2 Parts Factorization and Assembly Details

Part Definition. All structural MRIs are first registered to the common MNI152 space [1, 20, 29], after which anatomical labels are obtained using SynthSeg [3]. Based on these label maps, we define three fixed anatomical parts for compositional generation: the *left hemisphere*, the *right hemisphere*, and the *cerebellar-brainstem complex*. Specifically, the left hemisphere is formed from SynthSeg labels $\{2, 3, 4, 5, 10, 11, 12, 13, 17, 18, 26, 28\}$, the right hemisphere from $\{41, 42, 43,$ $44, 49, 50, 51, 52, 53, 54, 58, 60\}$, and the cerebellar-brainstem complex from $\{7, 8,$ $46, 47, 14, 15, 16\}$. The only explicit exclusion is CSF (label 24), which is not assigned to any part mask and is instead synthesized during the subsequent whole-brain refinement stage.

Part Cropping. Registration to the common MNI152 space allows us to use deterministic, template-aligned crops that generalize consistently across subjects. Within the $128 \times 128 \times 128$ whole-brain grid, the left and right hemisphere crops ($64 \times 128 \times 128$) are obtained by splitting the volume along the left-right axis (midline of the brain), while the cerebellar-brainstem crop

($128 \times 96 \times 64$) is taken from the inferior posterior portion of the registered brain. Because all samples share the same canonical coordinate system, these fixed crops can be mapped back unambiguously to their original locations when pasting refined parts into the whole-brain canvas during assembly to compose $\tilde{x}^{\mathrm{scaf}}$.

Part Masks. To obtain spatially consistent template masks, we segment the MNI152 template using SynthSeg and merge labels according to the three part definitions above, yielding one binary mask for the left hemisphere, one for the right hemisphere, and one for the cerebellar–brainstem complex. Each mask is dilated by 2 voxels ($3\,\mathrm{mm}$ at $1.5\,\mathrm{mm}$ isotropic resolution) to provide a small anatomical margin and to accommodate residual inter-subject variation after registration. These dilated masks, together with fixed bounding boxes, define the crops passed to the part models. Refined parts are pasted with the undilated masks intersected with the brain envelope; the cerebellar–brainstem mask is additionally eroded by 4 voxels, and each paste mask is feathered with a Gaussian of standard deviation 1.5 voxels. Inference needs no subject-specific segmentation or mask.

A.3 Structural Agreement with Real Test Volumes

To complement the distributional metrics, we compare SynthSeg-derived label maps of generated and real held-out test volumes for 17 anatomical regions. Dice measures volumetric overlap and HD95 the symmetric 95th-percentile Hausdorff surface distance in millimetres [45]. Compared with the LDM baseline, AnaDiffusion achieves higher Dice in 16 of 17 regions and lower HD95 in 16 of 17 regions (Table 3); the exceptions are the cerebral cortex for Dice and CSF for HD95. SynthSeg's training subjects are not public, so we cannot rule out that some held-out ADNI subjects contributed to its training labels. This applies to all SynthSeg-based results, including $|d|$ in Table 1.

Table 3: **Structural agreement between generated and real test volumes** using SynthSeg-derived anatomical labels. Dice (↑) and HD95 in mm (↓); the better value of each pair is in **bold**.

	LDM		Ours	
Region	Dice	HD95	Dice	HD95
Cerebral white matter	0.6461	3.88	**0.6657**	**3.61**
Cerebral cortex	**0.5475**	2.98	0.5353	**2.68**
Lateral ventricles	0.6941	5.88	**0.8127**	**3.44**
Inferior lateral ventricles	0.3257	4.45	**0.4497**	**3.37**
Cerebellar white matter	0.6764	3.95	**0.7516**	**3.03**
Cerebellar cortex	0.7457	4.01	**0.8031**	**2.79**
Thalamus	0.7500	3.49	**0.8245**	**2.49**
Caudate	0.5341	3.87	**0.6984**	**2.80**
Putamen	0.7043	3.10	**0.7868**	**2.25**
Pallidum	0.6165	3.28	**0.7460**	**2.21**
Third ventricle	0.6136	3.42	**0.7323**	**2.30**
Fourth ventricle	0.4588	3.83	**0.6524**	**2.52**
Hippocampus	0.6220	3.40	**0.7044**	**2.59**
Amygdala	0.6322	3.20	**0.7161**	**2.54**
CSF	0.4162	**4.10**	**0.4326**	4.25
Nucleus accumbens	0.4856	3.15	**0.6296**	**2.20**
Ventral diencephalon	0.6605	3.05	**0.7595**	**2.19**

A.4 ADNI Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

A.5 Broader Impacts

Synthetic brain MRIs with controllable regional anatomy could support data augmentation for underrepresented anatomical phenotypes, counterfactual studies of localized changes, and method development without sharing patient scans. They also carry risks. Synthetic scans could be mistaken for real data or used to fabricate evidence, models trained on patient data may memorize training subjects, and a model trained only on an older ADNI cohort may misrepresent other populations. AnaDiffusion is a research model that has not been validated for clinical use, and generated volumes should be labelled as synthetic.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the contributions (part-to-whole compositional latent diffusion, a shared hemisphere generator, localized part editing) and the evaluation scope (held-out subjects of a subject-disjoint ADNI split); Section 5 reports the supporting results with bootstrap confidence intervals.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 (Limitations) discusses the added training and inference complexity, the fixed anatomy-driven factorization, and the assumption of approximate correspondence with MNI152 space.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper makes no theoretical claims.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sections 3–4, Section 5.1 and Appendix A.1–A.2 give the architectures, noise schedule, injection timesteps, training settings, preprocessing, part definitions and data splits.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is released at the repository linked in the abstract. ADNI data are available to approved researchers under the ADNI data use agreement (Section 4).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 and Appendix A.1 give the optimizer, learning rates, epochs, optimizer steps, noise schedule and sampling settings; Section 4 gives the subject-level splits.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 1–2 and Figure 4 report means with 95% percentile confidence intervals from 5,000 subject-level cluster-bootstrap replicates (Section 5.1). No hypothesis tests are performed; overlapping intervals are noted in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A.1 reports the hardware. All models were trained on NVIDIA H200 GPUs with 141 GB of memory each.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses de-identified data from an existing study under its data use agreement and conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Appendix A.5 discusses positive uses (augmentation, counterfactual studies, less need to share patient scans) and risks (synthetic scans mistaken for real data, memorization of training subjects, bias from a single older cohort, unvalidated clinical use).

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We release code only. Trained model weights are not released because they were trained on ADNI data, and ADNI, as the data owner, governs the use and sharing of its data through its data use agreement. Researchers with ADNI access can train the models with the released code.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets are cited. ADNI data are used under the ADNI Data Use Agreement and are not redistributed. Code and pretrained models from prior work are used according to the terms of their public releases.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code is documented in the repository linked in the abstract.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or new research with human subjects; the study uses existing de-identified ADNI data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No new data were collected. ADNI participants were enrolled under IRB approval at the ADNI sites, and the data were accessed under the ADNI data use agreement.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the core methods; they were used only for editing (grammar, spelling, word choice), as declared at submission.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.