Yasinzai, M.N., Mito, R., & Pedersen, M. (2026). Wavelet-Fusion Diffusion Model for Multimodal Brain MRI Synthesis with Modality and Metadata Conditioning. *Imaging Neuroscience*, Advance Publication. https://doi.org/10.1162/IMAG.a.1458

Wavelet-Fusion Diffusion Model for Multimodal Brain MRI Synthesis with Modality and Metadata Conditioning

Muhammad Nabi Yasinzai,[1] Remika Mito,[2] Mangor Pedersen[1*]

[1]Department of Psychology & Neuroscience, Auckland University of Technology, Auckland, New Zealand.

[2]Department of Psychiatry, University of Melbourne, Melbourne, Victoria 3053, Australia.

[*]Correspondence: mangor.pedersen@aut.ac.nz

October 19, 2026

Keywords: brain MRI synthesis, latent diffusion, wavelet fusion, modality conditioning, metadata conditioning, BrainScape

Abstract: Multimodal MRI provides complementary information for neuroimaging analysis, with different modalities capturing distinct anatomical, tissue, and pathological features relevant to downstream artificial intelligence (AI) applications. Although large-scale structural MRI resources are increasingly available, modality coverage remains uneven across public and pooled neuroimaging datasets. Large pooled datasets span heterogeneous sites, scanners, and acquisition protocols and often contain incomplete or inconsistent demographic and clinical metadata. Synthetic MRI can supplement missing or underrepresented modalities and support metadata-conditioned synthetic cohorts. However, many existing approaches use limited modality combinations or relatively homogeneous datasets, limiting their applicability to heterogeneous pooled neuroimaging data. Diffusion models can generate high-fidelity, diverse MRI, but direct 3D voxel-space sampling is computationally expensive. Latent diffusion reduces this cost by generating in a compressed space, although information loss during compression can limit synthesis fidelity. We develop the Wavelet-Fusion Diffusion Model (WFDM), a 3D latent diffusion framework for multimodal brain MRI synthesis with target-modality and metadata conditioning. WFDM combines a Wavelet-Fusion variational

© 2026 The Authors. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

autoencoder (WF-VAE) compressor with a conditional 3D U-Net diffusion model trained in the latent space. Training and evaluation use 45,880 preprocessed T1w, T2w, T1Gd, and FLAIR scans from 157 BrainScape datasets. Compared with the BrainScape-trained MAISI (Medical AI for Synthetic Imaging) baseline, WFDM reduced pooled MedicalNet feature-space Fréchet distance (FID) from 0.00550 to 0.00404 and pooled Maximum Mean Discrepancy (MMD) from 0.1918 to 0.1471 while maintaining comparable sample diversity. WFDM achieved lower FID and MMD for T1w, T2w, and FLAIR, whereas MAISI achieved lower values for T1Gd. Rectified Flow sampling with 50 steps was 12.4–13.4 times faster than 1000-step DDPM sampling while achieving the lowest pooled FID and MMD in the sampler ablation. In T1w MRI, WFDM also more closely preserved age- and sex-associated morphometric patterns, including age-group morphometric separation, regional age–volume associations, and male–female differences. We release *BrainMint* for brain MRI synthesis workflows and *medmetric* for standardized evaluation of 3D medical image synthesis models.

1 Introduction

Multimodal structural MRI provides complementary information for neuroimaging analysis and clinical artificial intelligence (AI), because different MRI modalities capture distinct anatomical structures, tissue properties, and pathological features. For example, in clinical neuroradiology, multimodal MRI is particularly important for the diagnosis, treatment planning, and monitoring of brain tumors, where T1-weighted (T1w), T2-weighted (T2w), gadolinium-enhanced T1-weighted (T1Gd), and fluid-attenuated inversion recovery (FLAIR) MRI provide complementary information for identifying and delineating tumor subregions (Bakas et al., 2018; Menze et al., 2014). This dependence on multiple MRI contrasts makes modality availability an important consideration when developing and evaluating downstream AI models, as missing modalities can constrain downstream neuroimaging analyses and reduce the performance of AI models that rely on multimodal MRI inputs (Kim & Park, 2024; Meng et al., 2024). Acquiring the same set of multimodal MRI data for every subject is often impractical and clinically unnecessary, as MRI acquisition is shaped by diagnostic objectives and practical considerations, including scan time, cost, patient tolerance, and safety considerations (Jiang et al., 2023; Kim & Park, 2024). For example, gadolinium-based contrast is generally used only when clinically indicated (e.g., in brain tumor assessment (Zhou & Lu, 2013)), and rarely acquired in healthy and population cohorts, contributing to uneven modality availability across neuroimaging datasets.

Pooling datasets across studies can create larger neuroimaging resources, broaden demographic and

clinical representation, and increase the overall availability of different MRI modalities. However, pooling MRI data does not inherently resolve the issue of missing modalities. It can also introduce differences in scanner hardware, acquisition protocols, and preprocessing workflows, creating heterogeneity that affects the reliability of downstream neuroimaging analyses (Dadar & Duchesne, 2020; Yasinzai et al., 2025). This pattern of modality imbalance and acquisition heterogeneity is directly reflected in BrainScape, an open-source framework that we previously developed to aggregate and preprocess public anatomical MRI datasets (Yasinzai et al., 2025). BrainScape includes T1w, T2w, T1Gd, and FLAIR MRI from a large collection of public datasets while preserving source provenance, acquisition information, and available demographic and clinical metadata (Yasinzai et al., 2025).

Synthetic MRI generation can help address the underrepresentation of certain MRI modalities by augmenting the available data for these modalities with AI-generated images (Koetzier et al., 2024). Generative AI models can also be used for MRI modality translation, in which available MRI scans of a subject are used to synthesize missing contrasts, supporting downstream pipelines that rely on complete multimodal inputs (Jiang et al., 2023; Kim & Park, 2024; Meng et al., 2024). Metadata conditioning can further guide MRI synthesis according to attributes such as age and sex (Peng et al., 2024), but demographic and clinical variables are often sparse or inconsistently recorded across pooled datasets (Yasinzai et al., 2025).

Diffusion models have become attractive for medical image synthesis because of their strong sample fidelity and diversity, but direct 3D voxel-space diffusion is computationally expensive (Fan et al., 2024; Kazerouni et al., 2023). Previous work has reduced this burden through slice-based, cascaded, and latent-space MRI synthesis strategies (Friedrich et al., 2024; Zhu et al., 2023). However, patch-based and slice-based approaches may lose global 3D information, introduce inter-slice or volumetric inconsistencies, and produce artifacts that can limit downstream applicability (Friedrich et al., 2024; Kim & Park, 2024; Zhu et al., 2023). Latent diffusion reduces the memory and computational demands of high-resolution 3D MRI synthesis by using a learned compression model to encode 3D MRI into a lower-dimensional latent representation and performing diffusion in this compressed space (Guo et al., 2025; Kim & Park, 2024; Pinaya et al., 2022; Zhu et al., 2023). However, synthesis quality depends on the learned latent representation, as excessive compression can lead to information loss and degrade the quality of generated images (Friedrich et al., 2024; Rombach et al., 2022). Wavelet-domain diffusion applies the discrete wavelet transform (DWT) to represent 3D medical images as low-frequency and high-frequency subbands at reduced spatial resolution, performs diffusion denoising directly in this representation, and reconstructs the generated volumes using the inverse discrete wavelet transform (IDWT) (Friedrich et al., 2024). Beyond using wavelet coefficients as the direct synthesis domain, prior encoder–decoder models have introduced frequency information at multiple network resolutions. MWCNN introduces DWT and IDWT

operations with CNN blocks across the contracting and expanding subnetworks of a modified U-Net, enabling learned multilevel wavelet processing (Liu et al., 2018). WaveDiff introduces frequency-residual connections that convolutionally project DWT subbands into multiple encoder levels, whereas WaveFace applies iterative diffusion to the DWT low-frequency subband and recovers multilevel high-frequency subbands using a separate one-pass network (Miao et al., 2024; Phung et al., 2023).

Recent diffusion-based methods for subject-specific MRI modality completion and missing-modality synthesis use MRI from the same participant to synthesize missing modalities using multiple input modalities, a single source modality, or arbitrary configurations of available and missing modalities within a single model (Jiang et al., 2023; Kim & Park, 2024; Meng et al., 2024; Zhang et al., 2025). In contrast, cohort-level diffusion models have been developed to generate new 3D MRI volumes without conditioning on MRI scans from the same subject, operating either unconditionally or with demographic and metadata conditioning (Friedrich et al., 2024; Peng et al., 2024; Pinaya et al., 2022). Despite this progress, most MRI synthesis methods are developed and evaluated within individual datasets, and even models that support flexible source and target configurations remain constrained by the modality and metadata coverage available in those datasets (Meng et al., 2024; Zhang et al., 2025). Pooled neuroimaging resources combine heterogeneous datasets that differ in modality coverage, acquisition characteristics, and cohort composition, while demographic and clinical metadata can be sparse, inconsistently recorded, or unavailable across sources (Yasinzai et al., 2025). Models that assume a fixed set of MRI modalities or consistently available conditioning variables may therefore have limited applicability to pooled neuroimaging datasets, where both modality availability and metadata coverage vary across contributing sources (Lüpke et al., 2024; Yasinzai et al., 2025). Evaluation of synthetic MRI also remains difficult to standardize, as individual metrics capture different aspects of synthetic image quality and are sensitive to image distortions and normalization choices (Dohmen et al., 2025; Koetzier et al., 2024). For 3D MRI, metric values can depend on whether evaluation is performed slice-wise or on full 3D volumes. Distributional metrics can also depend on the feature representation used, complicating comparisons across studies.

Building on these gaps, we develop WFDM, a 3D latent diffusion pipeline for multimodal brain MRI synthesis designed for heterogeneous, pooled neuroimaging collections, supporting controllable generation of T1w, T2w, T1Gd, and FLAIR MRI through target-modality and metadata conditioning. WFDM combines a Wavelet-Fusion VAE compression stage with a conditional 3D U-Net diffusion model operating in the learned latent space, where generation is guided by the target modality and by demographic and clinical metadata, with an explicit metadata availability mask. We evaluate WFDM through latent-compressor and generator benchmarks, Wavelet Fusion Architecture and sampling ablations, inference-time efficiency analysis, and morphometric analyses assessing whole-brain and regional plausibility and

preservation of morphometric patterns associated with age and sex. To facilitate reuse, the software components developed for this study are released through *BrainMint*, an open-source Python package for brain MRI synthesis, compression, modality translation, inference, and evaluation (Yasinzai, 2026a).

2 Methodology

2.1 Ethics statement

This study is a secondary analysis of de-identified anatomical MRI data obtained from publicly available source datasets integrated through BrainScape. No new human data were collected for the present work. Ethical approval, participant consent, and requirements for data access are governed by the original contributing studies and repositories from which the BrainScape datasets were derived.

2.2 Dataset

2.2.1 BrainScape overview

The models are trained and evaluated on a fixed earlier BrainScape snapshot. The fully published BrainScape resource is a large-scale aggregation of 160 publicly available anatomical MRI datasets comprising 27,227 subjects and 46,583 scans after quality control. BrainScape provides four anatomical MRI contrasts used in this study, including T1w, T2w, T1Gd, and FLAIR, and retains demographic, clinical, and acquisition metadata whenever these fields are available in the originating datasets (Yasinzai et al., 2025).

For the present study, we used a fixed BrainScape snapshot comprising 157 datasets and 45,880 preprocessed MRI scans across T1w, T2w, T1Gd, and FLAIR. This snapshot was finalized before three additional datasets were added into the published BrainScape release. Therefore, the modality counts reported in Table 1 differ slightly from the published BrainScape statistics. The experimental snapshot contains fewer T1w and T2w scans, whereas the T1Gd and FLAIR counts are unchanged relative to the published BrainScape release. Table 1 summarizes the number of preprocessed BrainScape MRI volumes in the training, validation, and test splits used in this study. We use subject-aware splitting, ensuring that all scans from a given subject across modalities and sessions are assigned to a single split.

Table 1: BrainScape split summary.[a]

Modality	Total	Train	Val	Test
T1w	30,965	24,515	3,036	3,414
T2w	6,893	5,332	694	867
FLAIR	6,552	5,064	673	815
T1Gd	1,470	1,001	172	297
Total	45,880	35,912	4,575	5,393

[a] Counts correspond to the fixed earlier BrainScape snapshot used for WFDM training, validation, and testing in this study. This snapshot comprises 157 datasets and 45,880 MRI scans and was finalized before three additional datasets were incorporated into the published BrainScape release. The published BrainScape release contains 160 datasets and 46,583 MRI scans after quality control, including 31,411 T1w, 7,150 T2w, 1,470 T1Gd, and 6,552 FLAIR scans (Yasinzai et al., 2025).

2.2.2 BrainScape metadata conditioning

Each BrainScape dataset entry may include continuous, categorical, and binary demographic variables and clinical indicators. We use the available BrainScape demographic and clinical metadata to condition the generative models. The metadata conditioning schema defines a fixed order for 25 fields: age, height, weight, BMI, age group, sex, race, handedness, education, socioeconomic status, and clinical indicators for stroke, schizophrenia, depression, Attention-Deficit/Hyperactivity Disorder (ADHD), bipolar disorder, prosopagnosia, epilepsy, Focal Cortical Dysplasia (FCD), Hippocampal Sclerosis (HS), tumor, acute ischaemic stroke, Dysembryoplastic Neuroepithelial Tumor (DNT), Gliosis (GL), aneurysm, and Autism Spectrum Disorder (ASD). Using this schema, we deterministically transform the raw metadata associated with each BrainScape dataset entry into a fixed-length numerical vector and a corresponding metadata availability mask.

Formally, this metadata transformation is written as

$$(d_i, r_i) = \mathcal{T}_{\mathrm{meta}}(\mathrm{meta}_i)\,, \qquad d_i \in \mathbb{R}^P, \tag{1}$$

where i indexes a BrainScape dataset entry, meta_i denotes the raw demographic and clinical metadata record associated with entry i, $\mathcal{T}_{\mathrm{meta}}$ denotes a deterministic and non-learnable data-pipeline transformation, d_i is the resulting fixed-length numerical vector, r_i is the corresponding metadata availability mask, and $P = 25$ is the number of fields in the metadata schema.

The corresponding metadata availability mask satisfies

$$r_i \in \{0,1\}^P, \tag{2}$$

where $r_{i,j} = 1$ indicates that the jth metadata field is missing for entry i, whereas $r_{i,j} = 0$ indicates that it is available.

This transformation z-score normalizes the continuous variables age, height, weight, and BMI using statistics computed from the BrainScape dataset. Categorical variables are mapped to fixed integer identifiers, while binary demographic and clinical indicators are represented using three integer identifiers corresponding to *yes*, *no*, and *missing*. Missing continuous values are represented by zero after normalization. In contrast, missing categorical or clinical values are set to the designated *missing* integer identifier and, in both cases, the corresponding entry in r_i is set to 1. The resulting metadata vector d_i and availability mask r_i are then provided to the learnable metadata encoder described in Section 2.5.5, which produces the metadata-conditioning embedding used by the diffusion model.

2.3 Preprocessing

We preprocess the heterogeneous source datasets using the BrainScape BraTS preprocessing pipeline prior to model training (Yasinzai et al., 2025). In brief, the pipeline aligns modalities within each subject, standardizes volumes to atlas space, performs brain extraction and intensity normalization, and resamples all volumes to a uniform resolution.

When multiple modalities are available for a given entry, intra-subject co-registration is performed to a single reference modality selected by the following priority order: $\mathrm{T1w} > \mathrm{T2w} > \mathrm{T1Gd} > \mathrm{FLAIR}$. This step improves spatial correspondence among available modalities within each subject and reduces contrast misalignment before model training. Following co-registration, all volumes are rigidly registered to the SRI-24 atlas to standardize orientation and approximate anatomical scale (Yasinzai et al., 2025). Brain extraction is then performed using HD-BET to generate a brain mask, which is used to remove non-brain regions from each MRI volume. After registration and masking, the BrainScape BraTS pipeline applies intensity normalization and resamples all volumes to isotropic $1\,\mathrm{mm}$ spacing (Yasinzai et al., 2025).

For model training, volumes are represented in RAS orientation as channel-first tensors and cropped or padded to a fixed spatial size of $(H, W, D) = (192, 224, 160)$, giving tensor shape $(C, H, W, D) = (1, 192, 224, 160)$. These volumes undergo additional percentile-based intensity clipping followed by min-max scaling to $[0, 1]$.

2.4 Problem formulation

Let $\mathcal{M}$ denote the set of MRI contrasts considered in this study:

$$\mathcal{M} = \{\mathrm{T1w}, \mathrm{T2w}, \mathrm{T1Gd}, \mathrm{FLAIR}\}, \qquad |\mathcal{M}| = 4. \tag{3}$$

For subject entry i and modality $m \in \mathcal{M}$, we write the corresponding preprocessed single-channel 3D volume as

$$x_i^{(m)} \in \mathbb{R}^{H \times W \times D}, \qquad (H, W, D) = (192, 224, 160), \tag{4}$$

Because not every subject entry contains every modality, let $\mathcal{M}_i^{\mathrm{avail}} \subseteq \mathcal{M}$ denote the set of modalities available for subject entry i, and define the corresponding set of observed volumes as

$$X_i^{\mathrm{avail}} \triangleq \{x_i^{(m)} : m \in \mathcal{M}_i^{\mathrm{avail}}\}. \tag{5}$$

For training, each training example is indexed by a subject entry i and a selected target contrast τ, where $\tau \in \mathcal{M}_i^{\mathrm{avail}}$. The symbol τ denotes the requested output MRI contrast and $x_i^{(\tau)}$ corresponds to the target MRI volume. The label τ is provided to the model as conditioning, and the corresponding volume serves as the output MRI target ($y_{i,\tau}$):

$$y_{i,\tau} \triangleq x_i^{(\tau)}. \tag{6}$$

Using the metadata representation and metadata availability mask introduced in Section 2.2.2, the conditioning information associated with subject entry i and target contrast τ is written as

$$c_{i,\tau} = (\tau, d_i, r_i), \tag{7}$$

where τ specifies the desired output contrast, d_i denotes the metadata feature vector, and r_i is the corresponding metadata availability mask. The objective is to train a conditional generative model that learns the distribution of target MRI volumes given the conditioning information $c_{i,\tau}$.

2.5 Wavelet-Fusion Diffusion Model

The proposed Wavelet-Fusion Diffusion Model (WFDM) is a 3D latent diffusion framework conditioned on modality and metadata for generating a target MRI contrast volume $y_{i,\tau}$ in a compact latent space rather than directly in voxel space. Training is performed in two stages. In Stage A, a Wavelet-Fusion

VAE (WF-VAE) compressor, denoted by the encoder and decoder pair (E_ϕ, D_ψ), learns a compact latent representation for preprocessed brain MRI volumes. In Stage B, a conditional diffusion model is trained on latent representations produced by the selected frozen WF-VAE encoder and conditioned on the target modality, the metadata feature vector, and the corresponding metadata availability mask. After Stage A training is complete, the selected WF-VAE compressor is frozen. Real MRI volumes are encoded into latent tensors, the diffusion model is trained in this compressed space, and generated latent samples are mapped back to image space using the corresponding decoder. This two-stage design builds on prior latent diffusion models, with WFDM using the proposed WF-VAE as its latent compressor and explicitly conditioning the diffusion model on target modality and metadata (Guo et al., 2025; Pinaya et al., 2022; Rombach et al., 2022).

Figure 1 provides an overview of the two-stage WFDM framework. Panel A summarizes Stage A, the proposed WF-VAE compressor. The Wavelet Feature Extraction Module first applies a one-level 3D Haar DWT to the input MRI volume. The eight resulting Stage-1 subbands are concatenated channel-wise to form the Stage-1 coefficient stack $c^{(1)}$, which is projected to the Stage-1 wavelet feature representation $w^{(1)}$. The representation $w^{(1)}$ is used as input to the first encoder stage. The selected subset (LLL, LLH, LHL, HLL) is passed through a second one-level 3D Haar DWT, and the resulting Stage-2 subbands are concatenated to form $c^{(2)}$ and projected to $w^{(2)}$. The Stage-2 representation $w^{(2)}$ is concatenated channel-wise with the feature tensor produced after the first encoder downsampling stage. The rationale for the design of the Wavelet Feature Extraction Module is provided in Section 2.5.2. On the decoder side, the decoder backbone provides feature tensors at the two resolutions required by the Wavelet Reconstruction Module. The module projects these tensors into the predicted Stage-1 and Stage-2 coefficient stacks $\hat{c}^{(1)}$ and $\hat{c}^{(2)}$. The Stage-2 IDWT recovers the four selected Stage-1 subband tensors, which are added to the corresponding predicted Stage-1 coefficients before the final Stage-1 IDWT reconstructs the output volume. Panel B summarizes Stage B, latent diffusion in the learned WF-VAE latent space. During training, real MRI volumes are encoded into latent representations, which are progressively noised through the forward diffusion process. During inference, sampling begins from Gaussian noise, which is iteratively denoised in latent space before the resulting latent representation is decoded into MRI space.

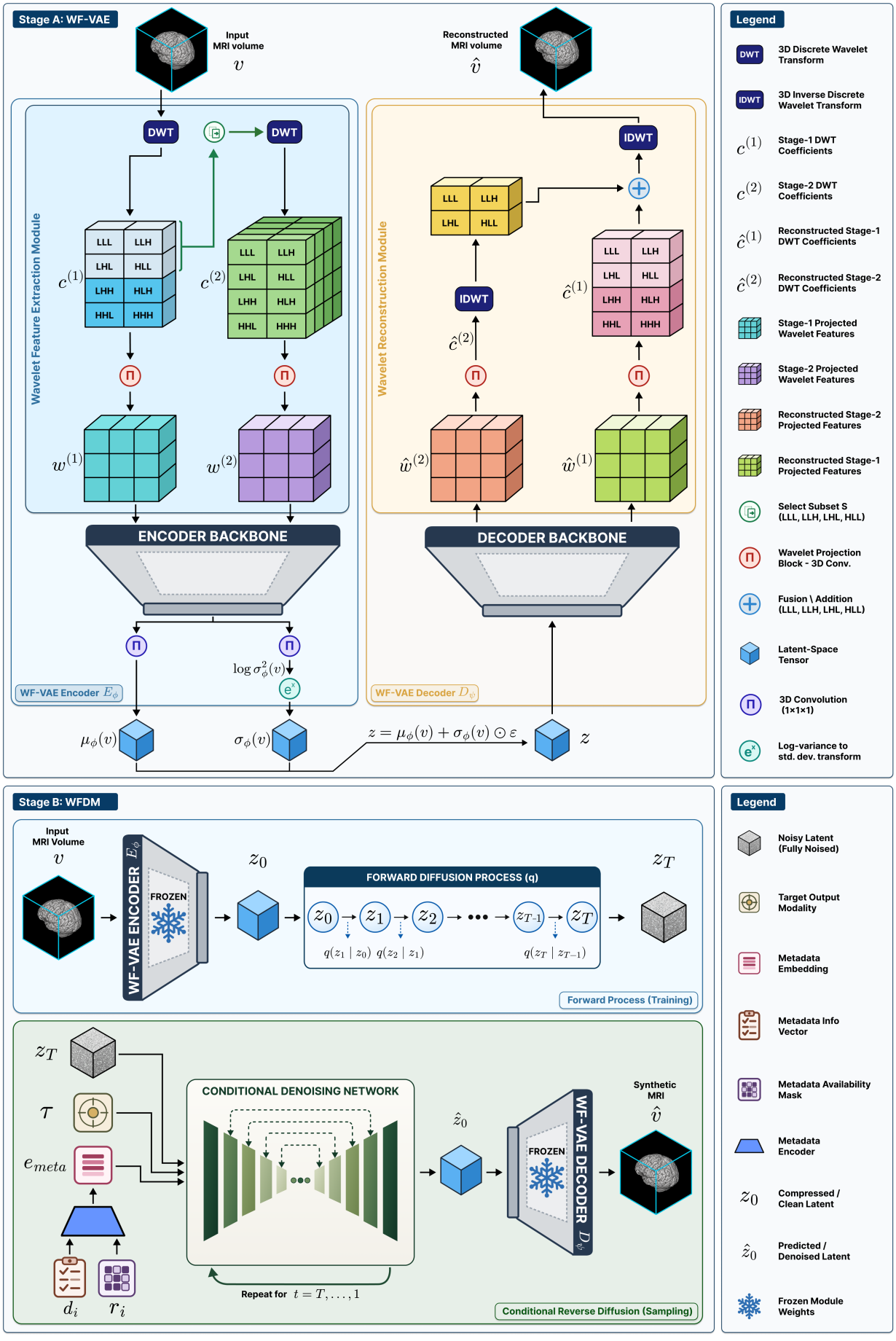

Stage A: WF-VAE
Input MRI volume
v
Reconstructed MRI volume
$\hat{v}$
DWT
DWT
IDWT
IDWT
Wavelet Feature Extraction Module
Wavelet Reconstruction Module
LLL
LLH
LHL
HLL
LHH
HLH
HHL
HHH
$c^{(1)}$
$c^{(2)}$
$\hat{c}^{(1)}$
$\hat{c}^{(2)}$
$w^{(1)}$
$w^{(2)}$
$\hat{w}^{(2)}$
$\hat{w}^{(1)}$
ENCODER BACKBONE
DECODER BACKBONE
$\log \sigma^2_\phi(v)$
WF-VAE Encoder E_ϕ
WF-VAE Decoder D_ψ
$\mu_\phi(v)$
$\sigma_\phi(v)$
$z = \mu_\phi(v) + \sigma_\phi(v) \odot \varepsilon$
z
Legend
3D Discrete Wavelet Transform
3D Inverse Discrete Wavelet Transform
Stage-1 DWT Coefficients
Stage-2 DWT Coefficients
Reconstructed Stage-1 DWT Coefficients
Reconstructed Stage-2 DWT Coefficients
Stage-1 Projected Wavelet Features
Stage-2 Projected Wavelet Features
Reconstructed Stage-2 Projected Features
Reconstructed Stage-1 Projected Features
Select Subset S (LLL, LLH, LHL, HLL)
Wavelet Projection Block - 3D Conv.
Fusion \ Addition (LLL, LLH, LHL, HLL)
Latent-Space Tensor
3D Convolution (1×1×1)
Log-variance to std. dev. transform
Stage B: WFDM
Input MRI Volume
v
WF-VAE ENCODER E_ϕ
FROZEN
z_0
FORWARD DIFFUSION PROCESS (q)
z_0 z_1 z_2 ••• z_{T-1} z_T
$q(z_1 \mid z_0)$ $q(z_2 \mid z_1)$
$q(z_T \mid z_{T-1})$
z_T
Forward Process (Training)
z_T
τ
e_{meta}
d_i
r_i
CONDITIONAL DENOISING NETWORK
Repeat for $t = T, \ldots, 1$
$\hat{z}_0$
WF-VAE DECODER D_ψ
FROZEN
Synthetic MRI
$\hat{v}$
Conditional Reverse Diffusion (Sampling)
Legend
Noisy Latent (Fully Noised)
Target Output Modality
Metadata Embedding
Metadata Info Vector
Metadata Availability Mask
Metadata Encoder
Compressed / Clean Latent
Predicted / Denoised Latent
Frozen Module Weights

Figure 1: Architecture of the proposed Wavelet-Fusion Diffusion Model (WFDM). **(A) Stage A: Wavelet-Fusion VAE (WF-VAE) compression.** Within the Wavelet Feature Extraction Module, a one-level 3D Haar discrete wavelet transform (DWT) is first applied to the input MRI volume, yielding eight Stage-1 subbands (Equation 21). The Stage-1 subbands are concatenated channel-wise to form $c^{(1)}$ and projected to the Stage-1 wavelet feature representation $w^{(1)}$ (Equation 23), which is used as the direct input to the first encoder stage. The selected subset (LLL, LLH, LHL, HLL) is transformed by a second DWT. The resulting Stage-2 subbands are concatenated to form $c^{(2)}$ and projected to $w^{(2)}$ (Equation 26). The complete encoder E_ϕ uses two separate $1 \times 1 \times 1$ convolutional layers to predict the posterior mean $\mu_\phi(v)$ and posterior log-variance, from which the posterior standard deviation $\sigma_\phi(v)$ is obtained before the latent z is sampled via the reparameterization trick (Equations 8–10). The complete decoder D_ψ reconstructs the output MRI volume from the latent z using the decoder backbone and the associated Wavelet Reconstruction Module (Equation 11). The decoder backbone provides feature tensors at the two resolutions required by the Wavelet Reconstruction Module, which projects them into the predicted coefficient stacks $\hat{c}^{(1)}$ and $\hat{c}^{(2)}$. The Stage-2 IDWT recovers the four selected Stage-1 subbands (LLL, LLH, LHL, HLL) from $\hat{c}^{(2)}$. These tensors are added to the corresponding tensors in $\hat{c}^{(1)}$ before the final Stage-1 IDWT reconstructs the MRI volume. The fixed perceptual feature extractor Φ and patch discriminator D_ω are auxiliary networks used only to compute the Stage-A perceptual and adversarial training objectives. Because these auxiliary networks are used only during Stage-A optimization and are not part of the WF-VAE encoder, decoder, or Stage-B WFDM, they are described in Section 2.5.1 and omitted from the architecture diagram to preserve visual clarity. **(B) Stage B: latent diffusion in the selected Wavelet-Fusion VAE space.** Real MRI volumes are encoded into latent tensors z_0, progressively noised during the forward diffusion process for training, and denoised by a conditional 3D U-Net during the reverse diffusion process. During inference, sampling begins from Gaussian noise $z_T \sim \mathcal{N}(0, I)$ and is denoised by the denoising network. The final denoised latent $\hat{z}_0$ is inverse-scaled and passed through the complete decoder of the WF-VAE to generate the synthetic MRI output.

2.5.1 Stage A: Wavelet-Fusion VAE

Three-dimensional structural MRI contains fine anatomical structures, including tissue boundaries and small-scale details, that can be degraded by compression. To better preserve anatomical information within a compact latent representation, we augment the autoencoder with the proposed Wavelet Fusion Architecture. It comprises a Wavelet Feature Extraction Module that supplies multilevel wavelet feature representations to the encoder backbone and a Wavelet Reconstruction Module that predicts the corresponding coefficient stacks from decoder-backbone output features and reconstructs the output MRI volume through inverse wavelet transforms.

Let $v \triangleq x_i^{(m)}$, with $m \in \mathcal{M}_i^{\text{avail}}$, denote a preprocessed single-modality input volume. The encoder E_ϕ

produces the parameters of a diagonal Gaussian approximate posterior over the latent representation:

$$q_\phi(z \mid v) = \mathcal{N}\big(\mu_\phi(v), \mathrm{diag}(\sigma_\phi^2(v))\big), \tag{8}$$

where z denotes the latent variable, $\mu_\phi(v)$ is the posterior mean tensor, and $\sigma_\phi^2(v)$ is the posterior variance tensor. In our implementation, $\mu_\phi(v)$ and $\log \sigma_\phi^2(v)$ are predicted by separate $1 \times 1 \times 1$ convolutional layers. After clamping the predicted log-variance tensor for numerical stability, the posterior standard deviation is obtained as

$$\sigma_\phi(v) = \exp\big(\tfrac{1}{2} \log \sigma_\phi^2(v)\big)\,. \tag{9}$$

A latent sample is then obtained using the reparameterization trick:

$$z = \mu_\phi(v) + \sigma_\phi(v) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I). \tag{10}$$

The decoder reconstructs the input as

$$\hat{v} = D_\psi(z), \tag{11}$$

where D_ψ denotes the complete WF-VAE decoder, which maps the unscaled autoencoder latent tensor to image space using the learned decoder backbone and the Wavelet Reconstruction Module described in Section 2.5.2. With a spatial downsampling factor of $f = 4$ and four latent channels, the latent tensor has shape

$$z \in \mathbb{R}^{4\times(H/4)\times(W/4)\times(D/4)} = \mathbb{R}^{4\times 48\times 56\times 40} \quad \text{for } (H, W, D) = (192, 224, 160). \tag{12}$$

The autoencoder is trained with a weighted combination of voxel-space reconstruction, perceptual similarity, Kullback-Leibler divergence (KL) regularization, and an adversarial objective based on a patch discriminator, following the general VAE-GAN compression strategy used in previous latent diffusion pipelines for medical imaging (Guo et al., 2025; Pinaya et al., 2022). The voxel-space reconstruction loss term is

$$\mathcal{L}_{\mathrm{recon}}(v, \hat{v}) = \|v - \hat{v}\|_1. \tag{13}$$

For the diagonal Gaussian posterior, the KL regularizer is

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}\big(q_\phi(z \mid v) \,\|\, \mathcal{N}(0, I)\big) = \frac{1}{2} \sum_k \Big(\mu_k^2 + \sigma_k^2 - \log(\sigma_k^2) - 1\Big), \tag{14}$$

where the sum runs over all latent elements (Kingma & Welling, 2013). To further improve perceptual

fidelity, we also include a feature reconstruction loss based on a fixed pretrained feature extractor $\Phi(\cdot)$. This encourages similarity in deep feature space beyond voxel-wise comparisons (Johnson et al., 2016). The perceptual feature reconstruction loss is

$$\mathcal{L}_{\text{perc}}(v, \hat{v}) = \big\|\Phi(v) - \Phi(\hat{v})\big\|_1. \tag{15}$$

The feature extractor Φ is fixed throughout Stage-A training and is used only to compute $\mathcal{L}_{\text{perc}}$. It is not part of E_ϕ, D_ψ, or the Stage-B WFDM architecture. We further introduce a patch-based discriminator D_ω during compression training, where ω denotes its trainable parameters. The discriminator is an auxiliary Stage-A training network and is not used during WF-VAE inference, Stage-B diffusion training, or WFDM sampling. For a real volume v and its reconstruction $\hat{v}$, the discriminator loss is

$$\mathcal{L}^D_{\text{adv}}(v, \hat{v}) = -\frac{1}{2}\Big(\mathbb{E}\big[\log D_\omega(v)\big] + \mathbb{E}\big[\log\big(1 - D_\omega(\hat{v})\big)\big]\Big), \tag{16}$$

and the generator-side adversarial term is

$$\mathcal{L}^G_{\text{adv}}(\hat{v}) = -\mathbb{E}\big[\log D_\omega(\hat{v})\big]. \tag{17}$$

This adversarial term encourages perceptually sharper and more realistic reconstructions (Goodfellow et al., 2020; Guo et al., 2025). The total autoencoder and discriminator objectives are

$$\mathcal{L}_{VAE} = \mathcal{L}_{\text{recon}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{adv}}\mathcal{L}^G_{\text{adv}}, \tag{18}$$

and

$$\mathcal{L}_D = \mathcal{L}^D_{\text{adv}}. \tag{19}$$

Training alternates between updating the autoencoder parameters (ϕ, ψ) by minimizing $\mathcal{L}_{VAE}$ with ω fixed, and updating the discriminator parameters ω by minimizing $\mathcal{L}_D$ with (ϕ, ψ) fixed (Guo et al., 2025; Pinaya et al., 2022).

2.5.2 Wavelet Fusion Architecture

The WF-VAE augments its encoder backbone with a Wavelet Feature Extraction Module and its decoder backbone with a Wavelet Reconstruction Module. Together with their connections to the learned encoder and decoder backbones, these modules form the proposed Wavelet Fusion Architecture illustrated in Figure 1. The Wavelet Feature Extraction Module produces Stage-1 and Stage-2 wavelet feature

representations at one-half and one-quarter of the input spatial resolution, respectively. The Stage-1 representation is the direct input to the first encoder stage, while the Stage-2 representation is concatenated with the feature tensor produced after that stage and its downsampling operation. The Wavelet Reconstruction Module projects decoder output features into predicted Stage-1 and Stage-2 coefficient stacks and reconstructs the output MRI volume.

Within the Wavelet Feature Extraction Module, each wavelet stage applies a one-level 3D Haar discrete wavelet transform (DWT) to decompose its input into low-frequency and high-frequency subbands along the three spatial dimensions. Throughout this section, we denote the 3D Haar discrete wavelet transform and its inverse by $\mathrm{DWT}(\cdot)$ and $\mathrm{IDWT}(\cdot)$, respectively. The DWT applies the separable low-pass and high-pass filters

$$\ell = \tfrac{1}{\sqrt{2}}[1, 1], \qquad h = \tfrac{1}{\sqrt{2}}[-1, 1], \tag{20}$$

along the three spatial axes and produces eight subbands, each at half the input resolution along each axis. This representation separates low-frequency structure from high-frequency detail while reducing the spatial resolution (Friedrich et al., 2024). A one-level DWT yields eight subbands:

$$\mathrm{DWT}(v) = \big(v^{LLL},\, v^{LLH},\, v^{LHL},\, v^{HLL},\, v^{LHH},\, v^{HLH},\, v^{HHL},\, v^{HHH}\big), \tag{21}$$

where each superscript records whether the low-pass or high-pass branch is applied along the corresponding spatial axis. Figure 2 visualizes the eight Stage-1 subbands. In the first wavelet stage, all eight subbands are concatenated channel-wise:

$$c^{(1)} = \mathrm{cat}\big(\,\mathrm{DWT}(v)\big). \tag{22}$$

The complete Stage-1 coefficient stack is mapped to a learned feature representation:

$$w^{(1)} = \Pi_1\big(c^{(1)}\big)\,, \tag{23}$$

Here, Π_1 denotes the trainable three-dimensional convolutional projection block that maps the Stage-1 coefficient stack $c^{(1)}$ to the Stage-1 wavelet feature representation $w^{(1)}$. The resulting representation $w^{(1)}$ is used as the direct input to the first encoder stage. Thus, the encoder backbone receives a trainable projection of the complete Stage-1 wavelet coefficient stack instead of the original voxel-space MRI volume. In the second wavelet stage, a selected subset of the Stage-1 subbands is transformed again:

$$c^{(2)} = \mathrm{cat}\Big(\,\mathrm{DWT}\,\big(S(c^{(1)})\big)\Big), \tag{24}$$

where $S(\cdot)$ denotes the subset selection operator and is defined as

$$S\big(c^{(1)}\big) = \big(v^{LLL},\, v^{LLH},\, v^{LHL},\, v^{HLL}\big). \tag{25}$$

The operator selects the all-low subband and the three single high-pass subbands in the stated order and returns them as the multichannel input to the Stage-2 DWT. As visualized in Figure 2, the selected subset combines the large-scale low-frequency anatomical structure represented by v^{LLL} with the single-axis directional detail captured by v^{LLH}, v^{LHL}, and v^{HLL}, each of which results from high-pass filtering along one spatial axis and low-pass filtering along the other two. The remaining subbands $(v^{LHH}, v^{HLH}, v^{HHL}, v^{HHH})$ combine high-pass responses along two or three spatial axes and encode finer multiaxial detail. This selection is motivated by the frequency-dependent grouping used by (Coupé et al., 2008), who retained the four lower-frequency subbands (LLL, LLH, LHL, HLL) from a lightly filtered MRI and combined them with the four higher-frequency subbands (LHH, HLH, HHL, HHH) from a more strongly filtered MRI to preserve the principal image structure while reducing high-frequency noise. We adopt this subband grouping to determine which Stage-1 subbands undergo the additional Stage-2 DWT. The remaining Stage-1 subbands $(v^{LHH}, v^{HLH}, v^{HHL}, v^{HHH})$ are not discarded: all eight Stage-1 subbands contribute to $w^{(1)}$, whereas $S(\cdot)$ determines only which four subbands undergo the additional decomposition. For an input with C channels, $S(c^{(1)})$ contains $4C$ channels. The Stage-2 DWT decomposes each of these channels into eight subbands, producing $8(4C) = 32C$ coefficient channels, which are concatenated to form the Stage-2 coefficient stack $c^{(2)}$. Applying the Stage-2 DWT to all eight Stage-1 subbands would instead decompose $8C$ input channels into $8(8C) = 64C$ coefficient channels. Thus, the selected subset provides Stage-2 DWT with the all-low subband and directional responses along each spatial axis while halving the width of $c^{(2)}$ relative to decomposing all eight Stage-1 subbands. The remaining multi-axis high-pass subbands are excluded only from the second decomposition and continue to contribute through $w^{(1)}$.

The Stage-2 coefficient stack is then projected as

$$w^{(2)} = \Pi_2\big(c^{(2)}\big)\,, \tag{26}$$

Here, Π_2 denotes the trainable three-dimensional convolutional projection block that maps the Stage-2 coefficient stack $c^{(2)}$ to the Stage-2 wavelet feature representation $w^{(2)}$. The DWT operations are fixed, whereas Π_1 and Π_2 are trainable. Consequently, $c^{(1)}$ and $c^{(2)}$ are deterministic coefficient stacks, while $w^{(1)}$ and $w^{(2)}$ are learned wavelet-derived feature representations.

The Stage-2 coefficient stack $c^{(2)}$ is computed from the selected Stage-1 subset $S(c^{(1)})$. The coefficient

stacks $c^{(1)}$ and $c^{(2)}$ remain distinct at one-half and one-quarter spatial resolution, respectively, and are projected separately into $w^{(1)}$ and $w^{(2)}$. These projected representations enter the encoder backbone at the corresponding spatial resolutions: $w^{(1)}$ is used as the direct input to the first encoder stage, while $w^{(2)}$ is concatenated channel-wise with the feature tensor produced after that stage and its downsampling operation. This alignment allows $w^{(2)}$ to be integrated without interpolation or additional learned spatial resampling. This design supplies a separately constructed wavelet-derived representation at the deeper encoder resolution, rather than relying only on the learned and downsampled features propagated from $w^{(1)}$. The one-quarter spatial resolution is retained in the WF-VAE latent representation defined in Equation 12.

For the separable 3D Haar transform, each Stage-1 coefficient is derived from a $2 \times 2 \times 2$ region of the input volume, whereas each Stage-2 coefficient is derived from a $4 \times 4 \times 4$ region. These sizes describe the fixed DWT coefficients before the trainable projection blocks. The second stage therefore provides a coarser spatial-frequency representation over a larger input region rather than duplicating the Stage-1 representation. This use of wavelet-derived representations at successive spatial scales is related to MWCNN, which interleaves wavelet transforms with CNN processing across successive U-Net resolutions, and to the frequency residual connections of WaveDiff, which concatenate and project DWT subbands for integration at multiple encoder levels (Liu et al., 2018; Phung et al., 2023). A single DWT level can provide an effective wavelet-domain representation, as demonstrated by WDM (Friedrich et al., 2024). In WF-VAE, the second stage provides an additional wavelet-derived representation aligned with the deeper one-quarter-resolution encoder stage.

On the decoder side, the decoder backbone provides feature tensors at the two resolutions required by the Wavelet Reconstruction Module. The Wavelet Reconstruction Module projects these tensors into the predicted Stage-1 and Stage-2 coefficient stacks $\hat{c}^{(1)}$ and $\hat{c}^{(2)}$. The Stage-2 IDWT maps $\hat{c}^{(2)}$ to the four selected Stage-1 subband tensors (LLL, LLH, LHL, HLL). These recovered subband tensors are added to the corresponding subband tensors in $\hat{c}^{(1)}$ to form $\tilde{c}^{(1)}$. The final Stage-1 IDWT then reconstructs the output MRI volume:

$$\hat{v} = \mathrm{IDWT}\big(\tilde{c}^{(1)}\big)\,. \tag{27}$$

2.5.3 Latent scaling

To improve the stability of diffusion training, we rescale the autoencoder latents using a single global factor estimated from the training set so that the latent tensors processed by the diffusion model have approximately unit variance. Let $\tilde{z}$ denote the raw autoencoder latent and let s_z denote the latent scaling

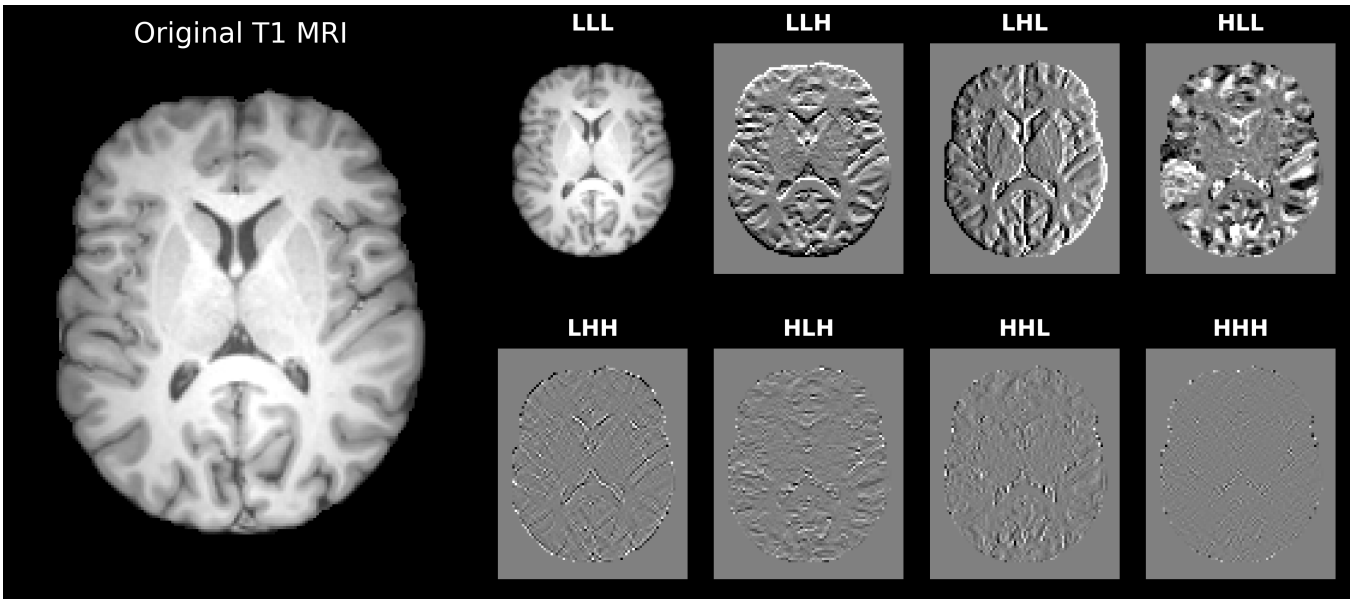


Figure 2: First-stage 3D Haar DWT decomposition of a T1w MRI volume from BrainScape. The original axial slice is from the AOMIC dataset (female; age 22.5 years) (Yasinzai et al., 2025). The top row shows the four lower-frequency subbands selected for the additional Stage-2 DWT, (LLL, LLH, LHL, HLL), comprising the all-low approximation subband and the three subbands containing one high-pass and two low-pass branches. The bottom row shows the remaining higher-frequency subbands, (LHH, HLH, HHL, HHH), which contain high-pass branches along two or three spatial axes. All eight Stage-1 subbands contribute to the Stage-1 wavelet feature representation $w^{(1)}$, whereas only the top-row lower-frequency subset undergoes the additional Stage-2 DWT. This frequency grouping follows the 3D MRI wavelet grouping used by (Coupé et al., 2008).

factor estimated from the training split. The scaled latent used by the diffusion model is

$$z = s_z \tilde{z}. \tag{28}$$

Throughout the remainder of this section, z denotes this scaled latent representation. The corresponding inverse scaling, $\tilde{z} = z/s_z$, is applied after the diffusion denoising process, immediately before decoding back to image space.

2.5.4 Stage B: Latent diffusion in Wavelet-Fusion VAE space

The latent diffusion stage is summarized schematically in Figure 1. In Stage B, we train a conditional denoising diffusion model in the latent space learned by the selected Wavelet-Fusion VAE compressor. Let z_0 denote the scaled clean latent representation of an input volume. The forward Markov process progressively adds Gaussian noise to z_0 over T diffusion steps, while a denoising network ϵ_θ is trained to predict the injected noise at arbitrary timesteps. During inference, sampling begins with Gaussian noise

$z_T \sim \mathcal{N}(0, I)$ and proceeds via iterative denoising to obtain a clean, scaled latent sample. The latent scaling is then reversed before the resulting unscaled latent is decoded back to image space (Guo et al., 2025). The forward noising process has a predefined variance schedule $\{\beta_t\}_{t=1}^{T}$ with

$$\alpha_t = 1 - \beta_t, \qquad \bar{\alpha}_t = \prod_{j=1}^{t} \alpha_j, \qquad \bar{\alpha}_0 = 1. \tag{29}$$

The Markov forward transition is

$$q(z_t \mid z_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}\, z_{t-1},\ (1-\alpha_t)I\big), \tag{30}$$

and the corresponding marginal distribution used to sample noisy latents during training is

$$q(z_t \mid z_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, z_0,\ (1-\bar{\alpha}_t)I\big), \qquad z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{31}$$

The denoising network ϵ_θ predicts the injected noise conditioned on $c_{i,\tau}$:

$$\hat{\epsilon} = \epsilon_\theta(z_t, t, c_{i,\tau}), \tag{32}$$

and is trained with the standard objective for predicting noise

$$\mathcal{L}_{\text{diff}} = \|\epsilon - \hat{\epsilon}\|_2^2. \tag{33}$$

At inference time, sampling starts from $z_T \sim \mathcal{N}(0, I)$ and follows the reverse process

$$p_\theta(z_{t-1} \mid z_t, c_{i,\tau}) = \mathcal{N}\big(\mu_\theta(z_t, t, c_{i,\tau}),\ \tilde{\beta}_t I\big), \qquad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\, \beta_t, \tag{34}$$

with mean parameterized by the predicted noise:

$$\mu_\theta(z_t, t, c_{i,\tau}) = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\, \epsilon_\theta(z_t, t, c_{i,\tau}) \right). \tag{35}$$

Iterating from $t = T$ to 1 yields a generated scaled clean latent estimate $\hat{z}_0$. Before decoding, inverse latent scaling is applied, and the generated MRI volume is obtained as

$$\hat{v} = D_\psi\left(\frac{\hat{z}_0}{s_z}\right), \tag{36}$$

where D_ψ denotes the complete frozen decoder of the selected Wavelet-Fusion VAE, as defined in

Equation 11. This decoder comprises the learned decoder backbone and the Wavelet Reconstruction Module detailed in Section 2.5.2, including the inverse wavelet reconstruction described in that section.

2.5.5 Modality and metadata conditioning

The latent diffusion denoiser is conditioned jointly on the target MRI modality and the associated demographic and clinical metadata.

Target-modality conditioning: The target modality τ is treated as a discrete class label indicating the desired output contrast. This label is provided as input to a learnable embedding layer within the 3D U-Net, and the resulting modality embedding is added element-wise to the timestep embedding.

Metadata conditioning and metadata availability encoding: The deterministic data-pipeline transformation described in Section 2.2.2 produces the metadata vector d_i and the corresponding metadata availability mask r_i for each BrainScape dataset entry i. These tensors are supplied to a dedicated learnable metadata encoder:

$$e_{\mathrm{meta},i} = \mathrm{ME}(d_i, r_i)\,, \qquad e_{\mathrm{meta},i} \in \mathbb{R}^{d_{\mathrm{meta}}}, \tag{37}$$

where ME denotes the learnable metadata encoder, $e_{\mathrm{meta},i}$ is the resulting metadata embedding for entry i, and d_{meta} denotes the fixed dimensionality of this embedding.

Within ME, the normalized continuous entries of d_i are retained as scalar inputs. For each categorical variable and clinical indicator, the corresponding integer-coded entry in d_i is passed through a separate learnable embedding layer. The continuous scalars and learned categorical and clinical embeddings are concatenated with the complete metadata availability mask r_i. The concatenated representation is then passed through two linear layers separated by a SiLU activation, with the second linear layer producing the metadata embedding $e_{\mathrm{meta},i}$. The inclusion of r_i provides ME with a separate indicator of metadata availability for each field, alongside the encoded values in d_i. Within the U-Net, $e_{\mathrm{meta},i}$ is further processed and concatenated with the combined timestep and modality embedding. The modality embedding layer and ME are optimized jointly with the denoising U-Net by minimizing the diffusion noise-prediction objective in Equation 33.

2.5.6 Training protocol

Training is performed in two stages. In Stage A, the autoencoder is trained on standardized 3D MRI volumes using the reconstruction, KL, perceptual, and adversarial objectives defined in Section 2.5.1. Optimization alternates between autoencoder and discriminator updates, together with learning-rate scheduling to stabilize convergence.

In Stage B, the trained autoencoder is frozen, while the denoising U-Net and the learnable modality and metadata conditioning components are optimized jointly using the corresponding scaled latent representations by minimizing the diffusion noise-prediction objective in Equation 33. To mitigate bias toward the most abundant target modalities, modality imbalance is addressed through weighted sampling, ensuring rarer target modalities are adequately represented during training.

2.6 Evaluation methodology

We assess latent compressor reconstruction fidelity, synthetic MRI distributional agreement and sample diversity, inference-time efficiency, morphometric plausibility, and preservation of metadata-associated regional morphometric patterns. This section defines the evaluation metrics and processing steps used across experiments. The corresponding experimental details are described in Section 3.

2.6.1 Latent compressor reconstruction fidelity

Latent compressors are evaluated using paired reconstruction metrics that compare each real volume v with its reconstruction $\hat{v}$. We report peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), multi-scale structural similarity index measure (MS-SSIM), learned perceptual image patch similarity (LPIPS), and mean squared error (MSE) in the preprocessed image space to quantify compression reconstruction fidelity.

2.6.2 Synthetic MRI distributional agreement

To evaluate the distributional agreement of the generated images, we use the *medmetric* Python package, which standardizes feature extraction and feature-space evaluation for 3D medical images using a fixed pretrained MedicalNet backbone (Yasinzai, 2026b). In our experiments, real and synthetic volumes are z-score normalized before feature extraction, consistent with the intended MedicalNet workflow for computing FID and MMD metrics.

Let F denote the fixed MedicalNet feature extractor. For each generated volume $\hat{v}_a$, $a = 1, \ldots, n_g$, and each real volume v_b, $b = 1, \ldots, n_r$, the corresponding feature vectors are obtained by passing both volumes through the same extractor:

$$g_a = F(\hat{v}_a) \in \mathbb{R}^d, \qquad r_b = F(v_b) \in \mathbb{R}^d. \tag{38}$$

Here, n_g and n_r denote the numbers of generated and real volumes, respectively, and d denotes the dimension of each MedicalNet feature vector. The corresponding generated and real feature sets are

$$G = \{g_a\}_{a=1}^{n_g}, \qquad R = \{r_b\}_{b=1}^{n_r}. \tag{39}$$

FID and MMD are computed from these feature sets in MedicalNet feature space rather than directly on voxel intensities. Let (μ_g, Σ_g) and (μ_r, Σ_r) denote their empirical means and covariance matrices, respectively. The Fréchet distance is defined as

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\Big(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\Big). \tag{40}$$

We additionally compute maximum mean discrepancy (MMD) using the same MedicalNet feature vectors. The unbiased estimator of squared MMD is

$$\widehat{\mathrm{MMD}}^2(G, R) = \frac{1}{n_g(n_g-1)} \sum_{\substack{a,a'=1 \\ a \neq a'}}^{n_g} k(g_a, g_{a'}) + \frac{1}{n_r(n_r-1)} \sum_{\substack{b,b'=1 \\ b \neq b'}}^{n_r} k(r_b, r_{b'}) - \frac{2}{n_g n_r} \sum_{a=1}^{n_g} \sum_{b=1}^{n_r} k(g_a, r_b). \tag{41}$$

The kernel is a Gaussian mixture:

$$k(x, y) = \sum_{\ell=1}^{L} w_\ell \exp\left(-\frac{\|x-y\|_2^2}{2\sigma_\ell^2}\right), \qquad w_\ell \geq 0, \quad \sum_{\ell=1}^{L} w_\ell = 1. \tag{42}$$

Finally, the reported MMD value is computed as

$$\mathrm{MMD}(G, R) = \sqrt{\max\Big(\widehat{\mathrm{MMD}}^2(G, R), 0\Big)}. \tag{43}$$

Lower FID and MMD indicate closer agreement between the generated and real feature distributions.

2.6.3 Synthetic MRI sample diversity

To assess diversity and detect mode collapse, we compute MS-SSIM between generated samples directly on generated volumes in image space. For each modality or conditioning bin, we sample K random pairs from N generated volumes and compute the average similarity between generated samples:

$$\text{MS-SSIM}_{\text{ff}} = \frac{1}{K}\sum_{k=1}^{K}\text{MS-SSIM}\big(\hat{v}^{(a_k)},\, \hat{v}^{(b_k)}\big)\,, \qquad a_k \neq b_k. \tag{44}$$

Lower values indicate lower self-similarity among generated samples and therefore greater sample diversity. All synthesis metrics are computed using a fixed evaluation protocol on the BrainScape test split. For WFDM and the BrainScape-trained MAISI internal baseline, generated samples are conditioned on target modalities and metadata drawn from the test split. The externally pretrained reference models are evaluated using output configurations for the modalities supported by those configurations. Results are reported by modality, together with pooled summaries where the supported modality sets permit pooling.

2.6.4 Synthetic MRI morphometric plausibility

SynthSeg regional morphometry and quality control: To perform morphometric comparisons between real and synthetic MRI, we use SynthSeg to obtain regional brain volumes, the estimated total intracranial volume (eTIV), and quality control (QC) scores (Billot et al., 2023). We apply the WASABI QC criterion to the SynthSeg outputs, retaining only segmentations with a general grey matter QC score of 0.7 or higher for morphometric analysis (Jafrasteh et al., 2025). Left and right hemisphere regional volumes are averaged to obtain bilateral volumetric measures, which are then normalized by eTIV (Jafrasteh et al., 2025). The normalized regional measure is defined as:

$$x_r = \frac{\left(V_r^L + V_r^R\right)/2}{\text{eTIV}}, \tag{45}$$

where eTIV denotes the estimated total intracranial volume, and V_r^L and V_r^R denote the corresponding left and right hemisphere volumes for region r, respectively. For the four unpaired volume measurements (the third ventricle, fourth ventricle, brainstem, and cerebrospinal fluid (CSF)), each volume is normalized by eTIV as $x_r = V_r/\text{eTIV}$, where V_r denotes the corresponding volume. The resulting 52-dimensional eTIV-normalized morphometric representation comprises 34 bilateral cortical measures, 14 bilateral global, subcortical, and ventricular volume measures, and four unpaired volumes. These eTIV-normalized measures are used in the morphometric analyses.

WASABI whole-brain morphometric distance: The Wasserstein-Based Anatomical Brain Index (WASABI) provides a single whole-brain measure of morphometric difference by comparing the multivariate distributions of the eTIV-normalized regional measures between two groups (Jafrasteh et al., 2025). Each MRI is represented by a 52-dimensional vector of eTIV-normalized regional measures, and WASABI calculates the difference between the resulting multivariate distributions using the squared Wasserstein distance. The resulting distance summarizes the overall morphometric difference across the 52 regional measures.

The compared groups can represent real and synthetic MRI datasets or population subgroups, such as different age or sex groups. A smaller WASABI distance indicates greater similarity between the morphometric distributions, whereas a larger distance indicates greater morphometric difference between the groups. For real and synthetic MRI comparisons, a lower distance indicates closer agreement between the synthetic and real morphometric distributions. For subgroup comparisons, we assess how closely the WASABI distance between synthetic subgroups matches the corresponding distance between real subgroups.

Regional morphometric differences: To assess regional morphometric differences between synthetic and real MRI, we compute signed Cohen's d from the eTIV-normalized regional volumes (Cohen, 2013; Peng et al., 2024). Cohen's d (d_{Cohen}) is the difference between two group means, scaled by their pooled standard deviation. In this study, we define the sign as synthetic minus real. A positive d_{Cohen} indicates a larger mean eTIV-normalized regional volume in the synthetic group than in the real group, while a negative value indicates a smaller mean volume. The magnitude $|d_{\text{Cohen}}|$ represents the size of the regional difference regardless of direction. Values closer to zero indicate smaller regional differences, while larger values indicate larger regional differences. For example, $|d_{\text{Cohen}}| = 1$ indicates that the two group means differ by one pooled standard deviation.

3 Experiments

3.1 Experiment A: Autoencoder benchmarking

The first experiment evaluated candidate latent compressors for 3D MRI synthesis. The goal was to identify a representation that preserved fine anatomical detail under compression while remaining practical for large-scale 3D synthesis. We compared latent compressors used in prior medical-image generation pipelines, including ALDM VQ-GAN (Kim & Park, 2024), BrainSynth VQ-VAE (Tudosiu et al., 2024),

the LDM-VAE (Pinaya et al., 2022), and the released MAISI VAE-GAN (Guo et al., 2025), all of which were pretrained on datasets other than BrainScape. For BrainSynth, we evaluated only the pretrained VQ-VAE component as an external reconstruction reference (Tudosiu et al., 2024).

We also evaluated autoencoders trained on BrainScape, including a MAISI VAE-GAN baseline, a Wavelet VAE, and the two proposed Wavelet-Fusion configurations: Wavelet-Fusion VAE (WF-VAE) and Wavelet-Fusion Self-Attention VAE (WF-SA-VAE). The Wavelet VAE applies a complete fixed two-stage wavelet decomposition to each input MRI volume, producing a 64-channel Stage-2 coefficient representation at one-quarter spatial resolution. This coefficient representation is used as the input to a simplified learned compression network that maps it to a 4-channel latent tensor. WF-VAE and WF-SA-VAE instead use a learned encoder–decoder backbone that incorporates the projected Stage-1 and Stage-2 wavelet representations at successive spatial resolutions. In the encoder, $w^{(1)}$ is used as the direct input to the first stage, while $w^{(2)}$ is concatenated with the downsampled feature tensor produced after that stage. The decoder predicts the Stage-1 and Stage-2 coefficient stacks and reconstructs the output through the inverse-wavelet operations described in Section 2.5.2. Experiment A therefore evaluates two different ways of incorporating the two-stage wavelet representation: using the complete Stage-2 coefficient representation as the input to a simplified compression network, as in Wavelet VAE, and integrating the projected Stage-1 and Stage-2 representations at successive spatial resolutions of a learned encoder–decoder backbone, as in WF-VAE and WF-SA-VAE. WF-SA-VAE differs from WF-VAE by including self-attention blocks in the deepest encoder and decoder stages.

In addition to the learned compressors, we evaluated the one-stage DWT used in WDM (Friedrich et al., 2024) and a complete two-stage DWT as fixed, invertible reconstruction references; neither transform requires training. For a single-channel input volume, the one-stage DWT produces eight wavelet subbands at one-half spatial resolution and contains $8(H/2)(W/2)(D/2) = HWD$ wavelet coefficients, while the complete two-stage DWT produces 64 wavelet subbands at one-quarter spatial resolution and contains $64(H/4)(W/4)(D/4) = HWD$ wavelet coefficients. The four-channel WF-SA-VAE latent at one-quarter spatial resolution instead contains $4(H/4)(W/4)(D/4) = HWD/16$ latent values, one-sixteenth as many values as either DWT reference. The DWT transforms were therefore used as fixed reconstruction references rather than as candidates for the compact learned latent used in Stage B. Using either DWT representation for generation would require training a separate diffusion model directly on the eight- or 64-channel wavelet-coefficient representation.

All compressors and fixed transforms were evaluated by reconstructing held-out 3D BrainScape volumes, with reconstruction quality assessed for each modality and for the pooled results using LPIPS, SSIM, MS-SSIM, PSNR, and MSE. The externally pretrained compressors were included as contextual references

evaluated on the same held-out BrainScape volumes, whereas compressor selection was restricted to the learned models trained on BrainScape in this study. The selected WF-SA-VAE was used as the latent compressor for the downstream diffusion and morphometric experiments.

3.2 Experiment B: Diffusion model benchmarking

The second experiment compared the proposed WFDM with a BrainScape-trained internal baseline and with externally pretrained reference generators for modalities where direct evaluation was possible. WFDM performs diffusion in the latent space of WF-SA-VAE and is conditioned on the target output modality together with a metadata embedding derived from demographic and clinical metadata and the corresponding metadata availability mask. The primary internal baseline MAISI is trained on the same BrainScape split using the same target-modality and metadata-conditioning interface as WFDM. WFDM and MAISI therefore provide the closest within-study comparison in Experiment B. We additionally evaluate externally pretrained generators including LDM (Pinaya et al., 2022), WDM (Friedrich et al., 2024), HA-GAN (Sun et al., 2022), and Med-DDPM (Dorjsembe et al., 2024) as contextual external references. These models retain their published training datasets, conditioning mechanisms, synthesis domains, and modality coverage; their results are therefore not interpreted as a controlled architecture comparison. Table 2 summarizes the evaluated generators, their conditioning strategies, synthesis domains, sampling configurations, training sources, and modality coverage. Evaluation of the benchmarked models was performed using the *medmetric* package (Yasinzai, 2026b). For distributional realism, FID and MMD were computed in the MedicalNet feature space. For diversity, MS-SSIM was computed between generated volumes in image space as a diversity proxy. Lower FID and MMD indicate closer alignment between real and synthetic feature distributions, whereas lower MS-SSIM between generated samples indicates lower self-similarity and therefore greater sample diversity.

3.3 Experiment C: Sampling and loss ablation

In the third experiment, we performed an ablation study on the proposed WFDM model family to evaluate how the generative formulation, noise schedule, prediction objective, and loss norm affect synthesis quality. Specifically, we compared DDPM-based sampling against Rectified Flow (RFlow) as alternative generative formulations. Within the DDPM formulation, we further ablated the noise schedule by comparing cosine and scaled-linear β schedules. For the prediction objective, DDPM variants used the v-prediction parameterization, while the RFlow variant used the corresponding velocity/flow prediction objective. Across selected configurations, we compared ℓ_1 and ℓ_2 regression losses to evaluate their

Table 2: Configuration and evaluation scope of the synthesis models in Experiment B.[a]

Model	Conditioning	Synthesis domain	Scheduler	Steps	Training source	Output modalities
WFDM	Modality + metadata	Latent (WF-SA-VAE)	RFlow	50	BrainScape	T1w, T2w, T1Gd, FLAIR
MAISI	Modality + metadata	Latent (MAISI VAE-GAN)	RFlow	50	BrainScape	T1w, T2w, T1Gd, FLAIR
LDM	UKB covariates[b]	Latent (LDM-VAE)	DDIM	50	UK Biobank	T1w
Med-DDPM	Segmentation mask	Pixel	DDPM	250	BraTS2021	T1w, T2w, T1Gd, FLAIR[c]
WDM	None	DWT coefficient space	DDPM	1000	BraTS2023	T1w
HA-GAN	None	Pixel	N/A (GAN)	N/A	GSP	T1w

[a] WFDM and MAISI were trained on the same BrainScape split using the same target-modality and metadata-conditioning interface. The remaining models are externally pretrained and evaluated using their released configurations for the modalities they support. Because these models differ in training data, conditioning, synthesis formulation, and modality coverage, they are included for broader comparison rather than as a controlled architecture benchmark.
[b] The published LDM is conditioned on age, sex, ventricular volume, and brain volume relative to the intracranial volume (Pinaya et al., 2022).
[c] The published multimodal Med-DDPM jointly generates all four BraTS modalities from a segmentation mask rather than generating a selected target modality (Dorjsembe et al., 2024).

impact. This experiment compared DDPM and RFlow sampling configurations to quantify how sampler choice affects synthesis quality and computational cost in latent diffusion. The default configuration for the final proposed model was selected by prioritizing lower pooled FID and MMD while verifying that MS-SSIM between generated samples remained competitive and did not indicate diversity collapse.

3.4 Experiment D: Inference-time efficiency benchmark

To assess the practical deployment cost of the evaluated generative models, we performed an inference-time efficiency benchmark on a single NVIDIA H100 GPU on New Zealand eScience Infrastructure (NeSI). All benchmarks were run on a single HPC node with 8 CPU cores and 64 GB host memory using CUDA 12.6.3, cuDNN 9.5.1.17, Python 3.11.13, and PyTorch 2.4.1. To isolate model execution performance, inference was performed with a batch size of 1, and the reported timing intervals measure only the model inference time, excluding disk I/O and data loading overhead.

For each benchmarked model, we first ran three warm-up generations to reduce initialization overhead and then recorded 30 timed generations. Timing was measured using synchronized GPU wall-clock timing, immediately before and after each inference call. For each configuration, we report the mean, median, and standard deviation of the time required to generate a single 3D MRI volume, together with the peak GPU memory usage and output tensor size.

This experiment is reported in two parts. The main efficiency benchmark compares all evaluated generators under the same measurement protocol. We also report a second efficiency table for the WFDM sampling and loss ablation study in Section 3.3, which quantifies how the generative formulation, noise

schedule, prediction objective, loss norm, and associated number of inference steps affect practical runtime within the same WFDM model family.

3.5 Experiment E: Wavelet Fusion Architecture ablation

The fifth experiment performed an ablation of the Wavelet Fusion Architecture in WF-SA-VAE. The experiment compared one-stage and two-stage wavelet configurations. Among the two-stage configurations, it also evaluated which subset of Stage-1 subbands was passed to the Stage-2 DWT. All configurations used the same self-attention VAE backbone architecture and training setup. The latent dimensions were the same across configurations, with four channels at one-quarter spatial resolution. For the proposed configuration, the WF-SA-VAE checkpoint from Experiment A was reused rather than retrained. All configurations were evaluated on the same held-out BrainScape volumes using the reconstruction protocol and metrics defined in Experiment A.

Four configurations were evaluated: Stage-1 only, Stage-2 LLL, Stage-2 Low4 (proposed), and Stage-2 All8. The Stage-1 only configuration removes the second wavelet stage, leaving $w^{(1)}$ as the only wavelet feature representation supplied to the encoder backbone. For the two-stage configurations, the selected Stage-1 subbands are decomposed by the Stage-2 DWT to form the Stage-2 coefficient stack $c^{(2)}$, which is then projected to the Stage-2 wavelet feature representation $w^{(2)}$. The Stage-2 LLL configuration applies the Stage-2 DWT only to the Stage-1 LLL subband, producing an $8C$-channel coefficient stack. The proposed WF-SA-VAE uses the Stage-2 Low4 configuration, which applies the Stage-2 DWT to (LLL, LLH, LHL, HLL) and produces a $32C$-channel coefficient stack, as described in Section 2.5.2. The Stage-2 All8 configuration applies the Stage-2 DWT to all eight Stage-1 subbands, producing a $64C$-channel coefficient stack.

3.6 Experiment F: Morphometric plausibility of synthetic MRI

The sixth experiment evaluated the morphometric plausibility of synthetic MRI using whole-brain, regional, and demographic subgroup analyses. All comparisons use held-out BrainScape test records and compare real MRI with synthetic MRI from WFDM and the BrainScape-trained MAISI baseline.

3.6.1 Whole-brain morphometric plausibility

Whole-brain morphometric plausibility is evaluated separately for T1w, T2w, and FLAIR using WASABI (Jafrasteh et al., 2025). Because WASABI has not been validated for MRI with lesions and tumors (Jafrasteh et al., 2025), and brain regions with tumors can produce unreliable SynthSeg segmentations (Huisman et al., 2024), we exclude MRIs with these pathologies from the primary morphometric analysis. Within BrainScape, we exclude all BraTS, ARCD, and SOOP records, as well as IDEAS records with positive DNT metadata and LSFD records with positive tumor metadata (Yasinzai et al., 2025). Because BraTS is the only BrainScape dataset providing T1Gd MRI, its exclusion also removes T1Gd from this analysis (Yasinzai et al., 2025).

For each source and modality, we report the SynthSeg QC pass rate, and WASABI is calculated only from scans satisfying the QC criterion defined in Section 2.6.4. For each modality, WASABI is calculated over 1000 repeated random subsampling iterations using $N = 100$ records per group for the Real–Real, Real–WFDM, and Real–MAISI comparisons. The Real–Real comparison provides a reference for the morphometric variability observed within the real population. The 1000 WASABI distances are summarized by their mean and standard deviation. For WFDM and MAISI, we additionally report the absolute difference between their mean WASABI distance and the corresponding Real–Real mean.

3.6.2 Demographic morphometric preservation

We use WASABI to evaluate whether synthetic T1w MRI preserves morphometric differences between demographic groups defined by age and sex, as observed in real MRI. This analysis uses T1w MRI from the test set after excluding scans with pathology. For each test record, all available demographic and clinical metadata, including age and sex, are used to condition WFDM and MAISI during synthetic MRI generation. Only matched Real, WFDM, and MAISI records satisfying the SynthSeg QC criterion defined in Section 2.6.4 are retained. Each demographic comparison further requires the corresponding age or sex information to be available.

We first compare two non-overlapping random groups from the same healthy population to provide a within-population reference. We then compare male and female groups and perform three age group comparisons: ages 5–19 versus 20–49, ages 20–49 versus 50–100, and ages 5–19 versus 50–100 years. For each comparison, the WASABI distance is calculated separately within Real, WFDM, and MAISI between the two demographic groups. Each comparison uses $N = 100$ records per group sampled without replacement and is repeated 1000 times. Within each iteration, the same selected test records are used across Real, WFDM, and MAISI. Greater similarity between the synthetic subgroup distance

distribution and the corresponding Real distribution indicates better morphometric preservation.

3.6.3 Regional morphometric differences

Regional morphometric differences were evaluated for T1w MRI using the 52 eTIV-normalized SynthSeg regional volume measures defined in Section 2.6.4. The analysis uses the same pathology exclusions and SynthSeg QC criterion as the whole-brain analysis in Section 3.6.1. Only T1w test records with QC-passing Real, WFDM, and MAISI MRIs were retained, ensuring that all three sources were compared using the same records. For each regional measure, the Real, WFDM, and MAISI volume distributions were visualized using boxplots. Regional differences were computed using Cohen's d for WFDM relative to Real and MAISI relative to Real, as defined in Section 2.6.4.

For anatomical visualization, the absolute Cohen's d values ($|d_{\mathrm{Cohen}}|$) were mapped onto the FreeSurfer `fsaverage` reference anatomy using MNE-Python (Fischl, 2012; Gramfort et al., 2013). The 34 bilateral cortical measures were displayed using the Desikan–Killiany `aparc` parcellation (Desikan et al., 2006). Subcortical, ventricular, and brainstem measures were displayed in two oblique three-dimensional views using the corresponding `aseg` labels (Fischl et al., 2002). The regional measurements were obtained from SynthSeg, while MNE-Python and the FreeSurfer reference anatomy were used only for anatomical visualization. For bilateral measures, the same $|d_{\mathrm{Cohen}}|$ value was assigned to the corresponding left and right hemisphere atlas labels.

3.7 Experiment G: Metadata-associated regional morphometry

The seventh experiment evaluated whether regional morphometric patterns associated with age and sex in Real T1w MRI are retained in synthetic T1w MRI generated using the corresponding metadata. The analysis uses the 52 eTIV-normalized SynthSeg regional volume measures defined in Section 2.6.4. All analyses use held-out T1w MRI from the BrainScape test set after applying the pathology exclusions described in Section 3.6.1. Only records for which the Real, WFDM, and MAISI MRIs all passed the SynthSeg QC criterion defined in Section 2.6.4 were retained. This leaves $2{,}894$ matched T1w test records before the additional selection criteria for each metadata analysis are applied. For each test sample, the WFDM and MAISI T1w images were generated using the same available demographic and clinical metadata associated with that record. For each analysis, one statistic is calculated for each of the 52 regional measures in Real, WFDM, and MAISI. For age, Pearson's r quantifies the association between age and eTIV-normalized regional volume. For sex, Cohen's d quantifies the difference in eTIV-normalized regional volume between males and females. For each synthetic source, the correspondence

between the synthetic and Real regional patterns is quantified using Pearson's correlation across the 52 regional statistics.

3.7.1 Age-associated regional morphometry

Test records with available age information from 18 to 80 years inclusive were retained, yielding $1{,}757$ matched records. For each regional measure and source, Pearson's r was calculated between age and eTIV-normalized regional volume. The 52 regional correlations from WFDM and MAISI were compared with the corresponding correlations from Real MRI. The age–volume relationships were also visualized for each region and source.

3.7.2 Sex-associated regional morphometry

Test records aged 18–80 years with available male or female sex labels were retained, yielding $1{,}733$ matched records, comprising 768 male and 965 female records. For each regional measure and source, Cohen's d was calculated between the male and female eTIV-normalized regional volumes, with the difference defined as male minus female. The 52 regional Cohen's d values from WFDM and MAISI were compared with the corresponding values from Real MRI. The male and female regional volume distributions were also visualized for each region and source.

4 Results

4.1 Compressor reconstruction results

Among the learned compressors trained on BrainScape, including the BrainScape-trained MAISI VAE-GAN baseline, WF-SA-VAE achieved the best pooled value for every reported reconstruction metric, with LPIPS of 0.0634, SSIM of 0.9646, MS-SSIM of 0.9947, PSNR of 30.79, and MSE of 0.000885 (Table 3). The released external MAISI VAE-GAN achieved the strongest absolute reconstruction performance among the learned compressors, but it was retained only as an external reference because it was trained on different source data. The fixed DWT rows were also not part of the candidate set because their near-lossless reconstruction retains all wavelet coefficients rather than forming a compact learned bottleneck. WF-SA-VAE was therefore selected as the latent compressor for the downstream diffusion and morphometric experiments.

Table 3: Pooled reconstruction metrics for learned compressors and fixed wavelet references.[a]

Model	Training dataset	Comp	Chans	LPIPS ↓	SSIM ↑	MS-SSIM ↑	PSNR ↑	MSE ↓
ALDM VQ-GAN (Kim & Park, 2024)	BraTS2021 / IXI	—	—	0.4739	0.7962	0.8243	16.86	0.506754
BrainSynth VQ-VAE (Tudosiu et al., 2024)	UKB / ADNI T1w MRI	—	—	0.6270	0.6631	0.8106	19.91	0.010698
LDM-VAE (Pinaya et al., 2022)	UK Biobank T1w MRI	1/8	3	0.2334	0.8641	0.9544	22.98	0.006548
MAISI VAE-GAN (BrainScape)	BrainScape	1/4	4	0.1418	0.6121	0.9828	27.32	0.002556
MAISI VAE-GAN (Guo et al., 2025)	Public CT+MRI	1/4	4	0.0415	0.9675	0.9953	32.55	0.000621
Wavelet VAE	BrainScape	1/4	4	0.2917	0.6273	0.9681	24.46	0.003785
WF-VAE	BrainScape	1/4	4	0.0798	0.9580	0.9927	29.65	0.001182
WF-SA-VAE	BrainScape	1/4	4	**0.0634**	**0.9646**	**0.9947**	**30.79**	**0.000885**
WDM-style DWT (1-stage) (Friedrich et al., 2024)	N/A	1/2	8	0.0000	1.0000	1.0000	82.63	0.000000
DWT reference (2-stage)	N/A	1/4	64	0.0000	1.0000	1.0000	76.61	0.000000

[a] The *Training dataset* column reports the source data used to train each learned compressor; N/A denotes fixed, non-learnable DWT references. The *Comp* column reports the spatial scale along each dimension; for example, $1/4$ indicates one-quarter of the input resolution. For the fixed DWT references, this spatial reduction does not reduce the total number of wavelet coefficients. The *Chans* column reports latent channels where applicable and wavelet-subband channels for the fixed DWT references. For VQ-based compressors, channel values are omitted because discrete codebook representations are not directly comparable with continuous VAE latents. Externally pretrained compressors are shown as contextual references evaluated on the same held-out BrainScape volumes; only the BrainScape-trained candidate compressors were considered for selection. LPIPS, SSIM, MS-SSIM, PSNR, and MSE compare input and reconstructed MRI volumes. Lower is better for LPIPS/MSE and higher is better for SSIM/MS-SSIM/PSNR. Bold indicates the selected compressor among the models trained on BrainScape. Externally pretrained compressors and fixed DWT references are excluded from this selection.

As shown in Table 3, under the same one-quarter spatial compression and four-channel latent dimensionality, the Wavelet VAE performed substantially worse than both Wavelet-Fusion variants across all reconstruction metrics. Under the evaluated configurations, these results support the proposed Wavelet Fusion Architecture over the Wavelet VAE configuration, which uses the complete 64-channel Stage-2 coefficient representation as the input to a simplified compression network.

Figure 3 provides a qualitative comparison of axial-slice reconstructions across the evaluated latent compressors. The figure includes four BrainScape test examples spanning different modalities, source datasets, demographic contexts, and available clinical labels, allowing reconstruction quality to be assessed under realistic cross-cohort heterogeneity (Yasinzai et al., 2025). The externally pretrained ALDM VQ-GAN (Kim & Park, 2024) shows poor reconstruction quality on the BrainScape examples. BrainSynth VQ-VAE (Tudosiu et al., 2024) shows prominent reconstruction artifacts and anatomical distortions, reflecting poor transfer of the pretrained VQ-VAE to BrainScape without retraining. Given this poor transfer, we retained BrainSynth as an external reconstruction reference in Experiment A and did not include it in the synthesis benchmark in Experiment B. LDM-VAE (Pinaya et al., 2022) produces blurrier reconstructions and shows poorer recovery of fine anatomical detail. The externally pretrained MAISI VAE-GAN (Guo et al., 2025) produces sharp and anatomically plausible reconstructions, consistent with its strong quantitative performance in Table 3. Among the compressors trained on BrainScape, Wavelet VAE reconstructions are visibly weaker, with blurrier structures and poorer recovery of fine anatomical

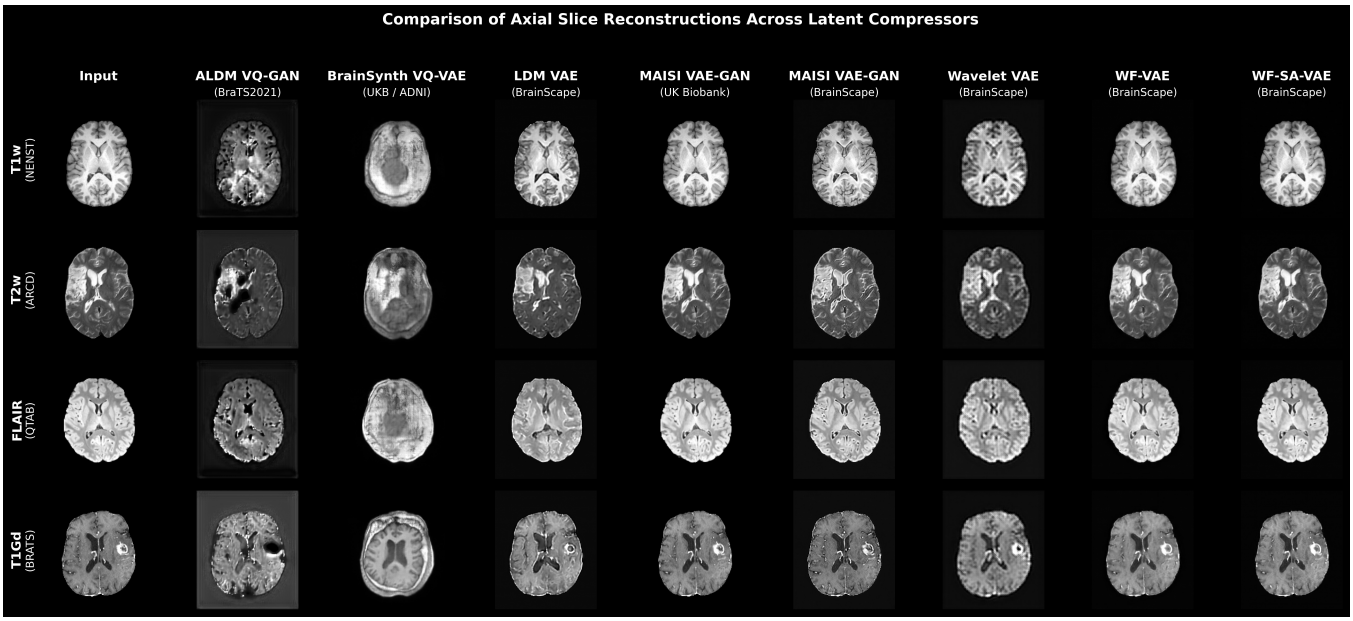


Figure 3: Qualitative comparison of axial-slice reconstructions across different latent compressors. Each row shows a different MRI modality drawn from a different BrainScape source dataset: FLAIR from QTAB (female, age 12 years), T1Gd from BraTS (tumor case), T1w from NENST (demographics unavailable), and T2w from ARCD (male, age 56 years, with a recorded stroke label) (Yasinzai et al., 2025). The first column shows the preprocessed input image and the remaining columns show reconstructions from the pretrained external compressors ALDM VQ-GAN (Kim & Park, 2024), BrainSynth VQ-VAE (Tudosiu et al., 2024), LDM-VAE (Pinaya et al., 2022), and the released MAISI VAE-GAN (Guo et al., 2025), together with the BrainScape-trained MAISI VAE-GAN, Wavelet VAE, and the two proposed Wavelet-Fusion variants, WF-VAE and WF-SA-VAE.

detail. The proposed Wavelet-Fusion variants preserve cortical boundaries, ventricular anatomy, and modality-specific contrast more accurately, with WF-SA-VAE producing the highest contrast and most anatomically coherent reconstructions within this group. The T1Gd example remains the most challenging case. Although the Wavelet-Fusion variants provide reasonable reconstructions, the externally pretrained MAISI VAE-GAN produces a cleaner and sharper T1Gd reconstruction. This is consistent with the modality-specific results showing that T1Gd is the weakest modality for our proposed latent compressor.

The full modality-specific reconstruction tables are provided in Supplementary Material Appendix A. Across T1w, T2w, T1Gd, and FLAIR, the results for each modality were consistent with the pooled comparison. In every modality, WF-SA-VAE remained the strongest learned compressor trained on BrainScape, while WF-VAE was the next strongest variant and the simpler Wavelet VAE lagged behind the variants using the Wavelet Fusion Architecture. The external MAISI VAE-GAN remained the strongest learned compressor overall, although the gap to WF-SA-VAE was comparatively small in T1w, T2w, and FLAIR and became more pronounced in T1Gd. T1Gd was the most challenging reconstruction case

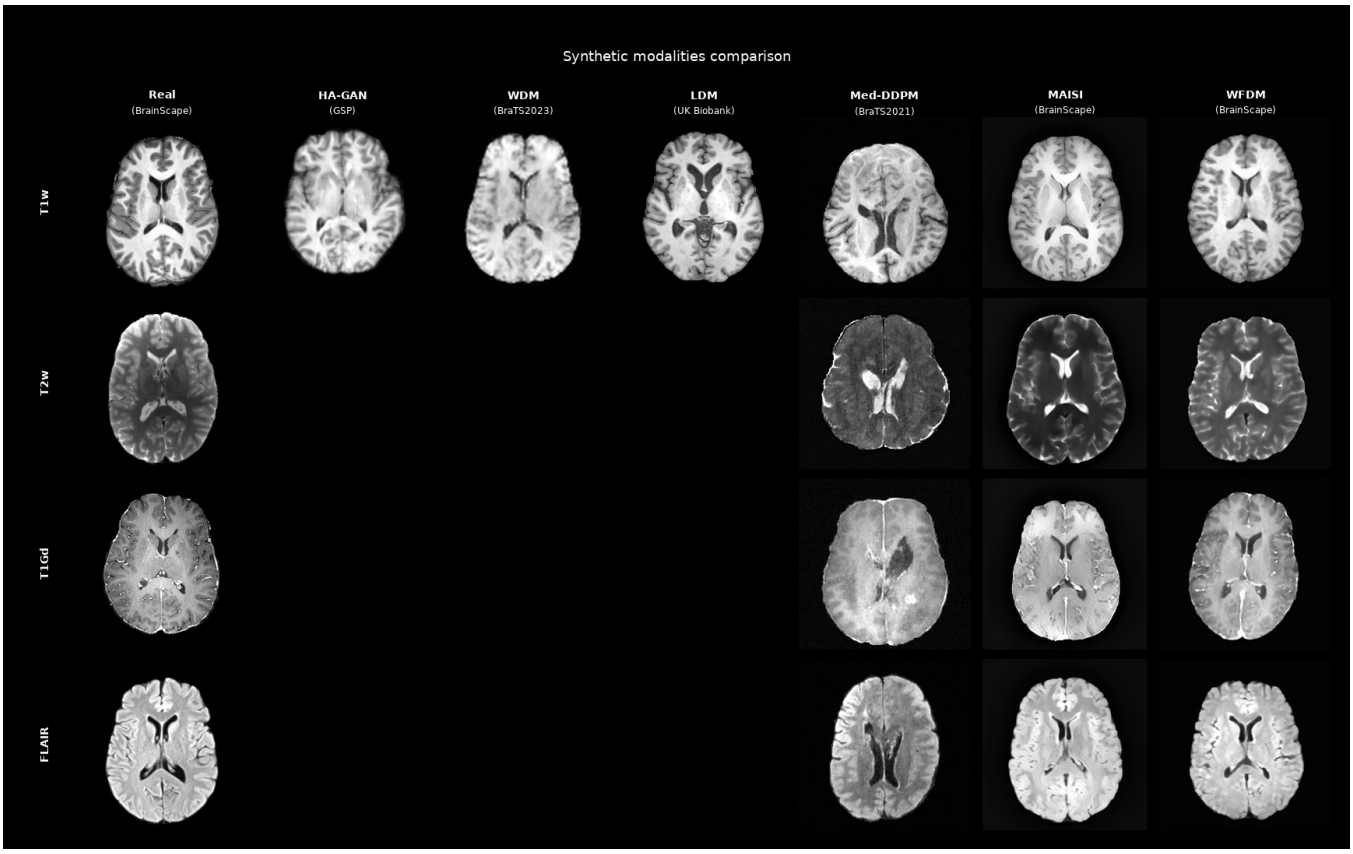


Figure 4: Qualitative axial-slice comparison across the four target modalities. Rows correspond to T1w, T2w, T1Gd, and FLAIR MRI modalities, and columns show samples from each generator. WFDM and the MAISI internal baseline form the matched BrainScape-trained comparison and use the same target-modality and metadata-conditioning interface. HA-GAN (Sun et al., 2022), WDM (Friedrich et al., 2024), LDM (Pinaya et al., 2022), and Med-DDPM (Dorjsembe et al., 2024) are externally pretrained references evaluated using their released configurations. Empty entries indicate that the corresponding external configuration does not support that target modality. Because the external models differ in training data, conditioning, synthesis formulation, and modality coverage, their columns provide contextual references rather than a controlled architecture comparison.

overall, generally showing lower SSIM and higher LPIPS and MSE across the learned models.

4.2 Diffusion model synthesis results

Figure 4 provides a qualitative axial-slice comparison across all four target modalities and also shows the differences in modality coverage among the evaluated synthetic MRI generators. The primary matched comparison is between WFDM and the MAISI internal baseline because both models were trained on BrainScape using the same target-modality and metadata-conditioning interface. The remaining columns provide contextual references rather than a controlled architecture comparison. Among these external

references, HA-GAN (Sun et al., 2022), WDM (Friedrich et al., 2024), and LDM (Pinaya et al., 2022) support T1w synthesis only in the evaluated configurations, which is reflected by the empty T2w, T1Gd, and FLAIR entries. The published multimodal Med-DDPM (Dorjsembe et al., 2024) generates all four modalities jointly, whereas MAISI and WFDM generate a requested target modality using modality and metadata conditioning. Under the evaluated configurations, the two BrainScape-trained latent diffusion models produce the most consistent anatomy and contrast appearance across the four modalities. Among the T1w-only external references, LDM is visually strongest, whereas HA-GAN and WDM are blurrier and recover less cortical detail. Med-DDPM samples more often exhibit texture or striping artifacts, reduced sharpness, and weaker structural fidelity than the BrainScape-trained models. Within the matched comparison, WFDM produces the stronger qualitative examples for T1w, T2w, and FLAIR, whereas the MAISI baseline produces a cleaner and sharper T1Gd example, consistent with the modality-specific results in Appendix B.

Table 4 summarizes the diffusion results under the evaluated configurations. Within the matched BrainScape-trained comparison, WFDM achieved lower pooled FID and MMD, with FID $= 0.00404259$ and MMD $= 0.147064$, than the MAISI internal baseline (Guo et al., 2025), with FID $= 0.00549863$ and MMD $= 0.191766$. This corresponds to approximately 26% lower pooled FID and 23% lower pooled MMD, while MS-SSIM diversity remained very similar between the two models. For the BrainScape-trained models and Med-DDPM, metrics are pooled across all four modalities, whereas metrics for the T1w-only external references are computed only on generated T1w samples. The limited modality coverage, absence of target-modality and metadata conditioning in most external models, and synthesis approaches that do not support flexible target-modality selection constitute task-specific architectural limitations for the multimodal synthesis task considered here. Consistent with these limitations, the external models generally achieved weaker distributional agreement in the present evaluation. Differences in training data and pretrained weights also contribute to the observed performance gaps.

The diffusion results for each modality are provided in Supplementary Material Appendix B. These results show that the performance of WFDM is contrast-dependent, with the strongest gains observed for T1w and T2w and smaller improvements for FLAIR. For T1w, WFDM reduced FID from 0.00335228 to 0.00159097 and MMD from 0.183375 to 0.123779 relative to the internal MAISI baseline. For T2w, WFDM improved FID from 0.0126402 to 0.00754393 and MMD from 0.266439 to 0.218096. FLAIR results were more closely matched, but WFDM still achieved the strongest FID and MMD. T1Gd remained the most challenging modality in the present analysis. For this contrast, the internal MAISI baseline achieved lower FID and MMD, whereas WFDM attained slightly lower MS-SSIM between generated samples. This weaker T1Gd synthesis performance is broadly consistent with the comparatively weaker T1Gd reconstruction performance of the selected Wavelet-Fusion latent compressor, which may

Table 4: Diffusion synthesis metrics for each evaluated generator.[a]

Model	Output modalities	FID ↓	MMD ↓	MS-SSIM ↓
WFDM	T1w, T2w, T1Gd, FLAIR	0.00404259	0.147064	0.754384
MAISI	T1w, T2w, T1Gd, FLAIR	0.00549863	0.191766	0.757653
LDM	T1w	0.01900610	0.411644	0.825484
Med-DDPM	T1w, T2w, T1Gd, FLAIR	0.05288810	0.341749	0.593099
WDM	T1w	0.05860420	0.444065	0.733082
HA-GAN	T1w	1.85993000	0.934294	0.694228

[a] For WFDM, MAISI, and Med-DDPM, metrics are pooled across T1w, T2w, T1Gd, and FLAIR. For LDM, WDM, and HA-GAN, metrics are computed on T1w samples only, reflecting the modality supported by their releases. FID and MMD measure distributional alignment in MedicalNet feature space, where lower values indicate closer agreement with the real BrainScape distribution. MS-SSIM is computed between generated samples as a diversity metric, where lower values indicate lower self-similarity and therefore greater sample diversity.

contribute to the remaining gap for this modality. Supplementary Material Appendix C provides qualitative comparisons across axial, sagittal, and coronal views for T1w, T2w, T1Gd, and FLAIR. These supplementary figures show that the strongest quantitative models also remain the most visually plausible across axial, sagittal, and coronal views, while again highlighting that T1Gd is the hardest contrast for the proposed WFDM generator.

4.3 Scheduler and sampling ablation

Table 5 reports the pooled solver and scheduler ablation study for the WFDM model family. RFlow with v-prediction and ℓ_2 loss achieved the strongest overall distributional agreement with real MRI, obtaining the lowest pooled FID and MMD while requiring only 50 inference steps. Relative to the 1000-step DDPM variants, RFlow with 50 inference steps reduced the number of sampling steps by a factor of 20 while also improving the main distributional agreement metrics (FID and MMD). Within DDPM, the cosine schedule consistently outperformed the scaled-linear alternative on FID and MMD. Although the scaled-linear schedule produced the lowest MS-SSIM among generated samples, this diversity gain was accompanied by substantially worse FID and MMD scores. Lower self-similarity alone is insufficient evidence of better generation quality when distributional alignment with real images deteriorates. Overall, RFlow provided the best balance between distributional agreement, diversity, and sampling efficiency, and was therefore used as the default WFDM sampler.

Table 5: Overall scheduler and sampler comparison for the WFDM model family.

Scheduler / Sampler	Pred. type	Loss	Inference steps	Overall FID ↓	Overall MMD ↓	Overall MS-SSIM ↓
DDPM (cosine β schedule)	v	ℓ_1	1000	0.00535477	0.184523	0.761192
DDPM (cosine β schedule)	v	ℓ_2	1000	0.00567393	0.163462	0.742717
DDPM (scaled-linear β schedule)	v	ℓ_2	1000	0.05349090	0.337763	0.436904
RFlow	v	ℓ_2	50	0.00404259	0.147064	0.754384

4.4 Inference-time efficiency results

Table 6 summarizes the inference-time benchmark across all evaluated generators. The primary metric is latency per generated 3D MRI volume, measured around the model inference call and excluding data loading and file writing. Peak GPU memory is reported to reflect the practical resource requirements of each model under the same deployment setting.

Across the evaluated diffusion models, runtime varies with the number of denoising steps, synthesis domain, model architecture, and output size. WFDM generated a full $(192, 224, 160)$ volume in 2.161 seconds on average using 50 RFlow steps, compared with 1.919 seconds for the MAISI baseline. The T1w-only LDM reference required 7.453 seconds, WDM required 24.128 seconds for a smaller $(128, 128, 128)$ output, and Med-DDPM required 95.86 seconds for a $(192, 192, 144)$ MRI output. HA-GAN was the fastest model at 0.034 seconds per sample because it does not require iterative denoising, but its FID and MMD were substantially higher than those of the latent diffusion models trained on BrainScape.

Table 7 reports the corresponding efficiency benchmark for the WFDM sampler ablation study. This table isolates the effect of replacing 1000-step DDPM inference with 50-step RFlow inference while holding the remaining WFDM model family fixed. The three DDPM configurations required 26.849 to 28.875 seconds per generated volume on average, whereas RFlow required only 2.161 seconds. This makes RFlow 12.4 to 13.4 times faster than the DDPM configurations, while also achieving the lowest pooled FID and MMD in Table 5. Together, the quality and runtime results indicate that RFlow is the most practical WFDM sampling configuration, combining the strongest distributional alignment with substantially faster inference.

4.5 Wavelet-fusion architecture ablation

Table 8 reports the pooled reconstruction results for the four Wavelet Fusion Architecture configurations. Stage-2 Low4 achieved the lowest LPIPS (0.0634), which was 7.6% lower than Stage-2 LLL (0.0686) and 22.1% lower than Stage-2 All8 (0.0814). Stage-2 LLL achieved the highest SSIM (0.9681), compared with 0.9646 for Stage-2 Low4 and 0.9572 for Stage-2 All8. Among the two-stage configurations, Stage-

Table 6: Inference-time benchmark of evaluated 3D synthetic MRI generators.[a]

Model	Steps	Output size	Mean s/vol ↓	Median s/vol ↓	Std s/vol ↓	Peak mem. GB ↓
WFDM	50	$(192, 224, 160)$	2.161	2.106	0.0871	3.002
MAISI	50	$(192, 224, 160)$	1.919	1.939	0.0855	14.370
LDM	50	$(192, 224, 160)$	7.453	7.453	0.0241	13.470
WDM	1000	$(128, 128, 128)$	24.128	24.183	0.2439	0.865
Med-DDPM	250	$(192, 192, 144)$	95.860	95.890	0.1193	13.320
HA-GAN	N/A	$(192, 224, 160)$	0.034	0.034	0.0004	4.769

[a] Inference-time benchmark performed on a single NVIDIA H100 GPU on NeSI with batch size 1. Latency is reported in seconds per generated volume and excludes data loading and file writing. Each configuration was measured over 30 timed runs after 3 warm-up runs. *Steps* denotes the number of iterative sampling steps. N/A indicates a non-iterative GAN-based generator. *Output size* reports the generated 3D volume dimensions, and *Peak mem.* reports peak GPU memory usage in GB. Lower values indicate faster inference or lower memory use.

Table 7: Inference-time benchmark of WFDM under different sampler settings.[a]

Sampler	Pred. type	Loss	Steps	Mean s/vol ↓	Median s/vol ↓	Std s/vol ↓	Peak mem. GB ↓
DDPM (cosine β)	v	ℓ_1	1000	28.875	28.844	1.2251	3.002
DDPM (cosine β)	v	ℓ_2	1000	28.194	28.448	1.4040	3.002
DDPM (scaled-linear β)	v	ℓ_2	1000	26.849	26.603	0.5813	3.002
RFlow	v	ℓ_2	50	2.161	2.106	0.0871	3.002

[a] WFDM sampler settings are benchmarked using the same protocol as Table 6, with output sizes $(192, 224, 160)$. *Pred. type*, *Loss*, and *Steps* denote the prediction target, training loss objective, and number of inference sampling steps, respectively. Lower values indicate faster inference or lower peak GPU memory use.

2 LLL also achieved the highest MS-SSIM and PSNR and the lowest MSE. These results show a clear trade-off between the two strongest two-stage configurations: Stage-2 Low4 achieved the best LPIPS, while Stage-2 LLL performed better on SSIM, MS-SSIM, PSNR, and MSE.

Stage-1 only configuration achieved the highest overall MS-SSIM (0.9951), PSNR (31.77) and the lowest MSE (0.000717). Stage-2 LLL was a close second on these metrics, with an MS-SSIM of 0.9950, PSNR of 31.36, and MSE of 0.000777. However, Stage-1 only had lower SSIM and substantially higher LPIPS than both Stage-2 LLL and Stage-2 Low4. Stage-2 All8 did not achieve the best pooled value for any metric. Stage-2 All8 therefore provided no consistent reconstruction advantage over the more selective two-stage configurations. The modality-specific results in Supplementary Material Appendix A show a similar pattern across T1w, T2w, T1Gd, and FLAIR. Stage-2 Low4 achieved the lowest LPIPS for T1w, T1Gd, and FLAIR, while Stage-2 LLL achieved the lowest LPIPS for T2w and the highest SSIM for all

Table 8: Pooled reconstruction results for the WF-SA-VAE wavelet-fusion architecture ablation.[a]

Variant	Stages	Stage-2 input	Comp	LPIPS ↓	SSIM ↑	MS-SSIM ↑	PSNR ↑	MSE ↓
WF-SA-VAE (proposed)	2	Low4	1/4	**0.0634**	0.9646	0.9947	30.79	0.000885
WF-SA-VAE (Stage-1 only)	1	—	1/4	0.1045	0.9475	**0.9951**	**31.77**	**0.000717**
WF-SA-VAE (Stage-2 LLL)	2	LLL	1/4	0.0686	**0.9681**	0.9950	31.36	0.000777
WF-SA-VAE (All8)	2	All8	1/4	0.0814	0.9572	0.9942	30.90	0.000862

[a] Stage-2 Low4 is our proposed WF-SA-VAE compressor. The Stage-1 only WF-SA-VAE configuration omits the second-stage DWT decomposition. WF-SA-VAE (Stage-2 LLL) applies the Stage-2 DWT only to the Stage-1 LLL subband, while WF-SA-VAE (Stage-2 All8) applies it to all eight Stage-1 subbands.

four modalities.

4.6 Morphometric plausibility results

4.6.1 Whole-brain morphometric plausibility

Table 9 summarizes SynthSeg QC performance and whole-brain WASABI distances for T1w, T2w, and FLAIR. SynthSeg QC pass rates were high for T1w and T2w across all three sources, ranging from 99.5–99.8% and 98.9–99.4%, respectively. FLAIR showed substantially lower QC pass rates, with 73.5% for Real, 44.1% for WFDM, and 29.7% for MAISI. Because WASABI was calculated only from QC-passing MRIs, the FLAIR comparison reflects this selected subset of the test data. Among MRIs that passed QC, WFDM was closer to the Real–Real morphometric reference than MAISI across all three modalities. For T1w, the WFDM WASABI distance remained close to the Real–Real reference, whereas the MAISI distance was larger. For T2w and FLAIR, WFDM remained closer to the Real–Real reference than MAISI, but its absolute difference from the corresponding Real–Real mean WASABI distance was larger than for T1w. Overall, WFDM showed closer morphometric agreement with Real MRI than MAISI across all three evaluated modalities.

4.6.2 Demographic subgroup morphometric preservation

Figure 5 shows the repeated subsampling WASABI distributions for the demographic comparisons. For readability, WASABI distances in this subsection are reported in units of 10^{-3}, consistent with the scale used in the figure. The mean same-population distances were 0.0818 for Real, 0.0458 for WFDM, and 0.0365 for MAISI. Both synthetic sources showed lower same-population distances than Real, with WFDM closer to the Real reference than MAISI. For the male–female comparison, the corresponding

Table 9: SynthSeg WASABI whole-brain morphometric distances for real and synthetic MRI.[a]

Modality	Source	QC pass n/N (%) ↑	WASABI D_{W}^2 mean ± SD ($\times 10^{-3}$) ↓	$\lvert\Delta_{\mathrm{Real-Real}}\rvert$ ($\times 10^{-3}$) ↓
T1w	Real	2912/2927 (99.5)	0.084 ± 0.019	0.000
	WFDM	2921/2927 (99.8)	0.095 ± 0.029	0.011
	MAISI	2912/2927 (99.5)	0.374 ± 0.045	0.290
T2w	Real	532/536 (99.3)	0.078 ± 0.015	0.000
	WFDM	530/536 (98.9)	0.239 ± 0.032	0.161
	MAISI	533/536 (99.4)	0.337 ± 0.036	0.259
FLAIR	Real	255/347 (73.5)	0.080 ± 0.013	0.000
	WFDM	153/347 (44.1)	0.202 ± 0.037	0.122
	MAISI	103/347 (29.7)	0.376 ± 0.045	0.296

[a] QC pass n/N reports the number of MRIs passing SynthSeg QC (n) relative to the total number of eligible MRIs evaluated (N), with the corresponding percentage in parentheses. The Real row reports the Real–Real WASABI reference, while the WFDM and MAISI rows report the Real–WFDM and Real–MAISI comparisons, respectively. WASABI D_{W}^2 is reported as the mean ± SD across repeated subsamples in units of 10^{-3}. The final column reports the absolute difference between the mean Real–Synthetic WASABI distance and the corresponding Real–Real mean, also in units of 10^{-3}.

mean distances were 0.0874, 0.0535, and 0.0443 for Real, WFDM, and MAISI, respectively. These values were slightly higher than the corresponding same-population distances, while WFDM remained closer to the Real value than MAISI.

The age-group comparisons showed substantially larger morphometric separation and the same ordering across Real, WFDM, and MAISI. For ages 5–19 versus 20–49 years, the mean D_{W}^2 values were 0.277, 0.194, and 0.0902 for Real, WFDM, and MAISI, respectively. For ages 20–49 versus 50–100 years, the corresponding values were 1.32, 0.939, and 0.452; for ages 5–19 versus 50–100 years, they were 2.25, 1.55, and 0.761.

$$D_{\mathrm{W}}^2(5\text{–}19,\ 20\text{–}49) < D_{\mathrm{W}}^2(20\text{–}49,\ 50\text{–}100) < D_{\mathrm{W}}^2(5\text{–}19,\ 50\text{–}100).$$

Across the three age contrasts, the WFDM mean distance was 68.8–71.1% of the corresponding Real value, compared with 32.5–34.2% for MAISI. Thus, both synthetic models preserved the ordering of the age-group differences, while WFDM more closely preserved their magnitude.

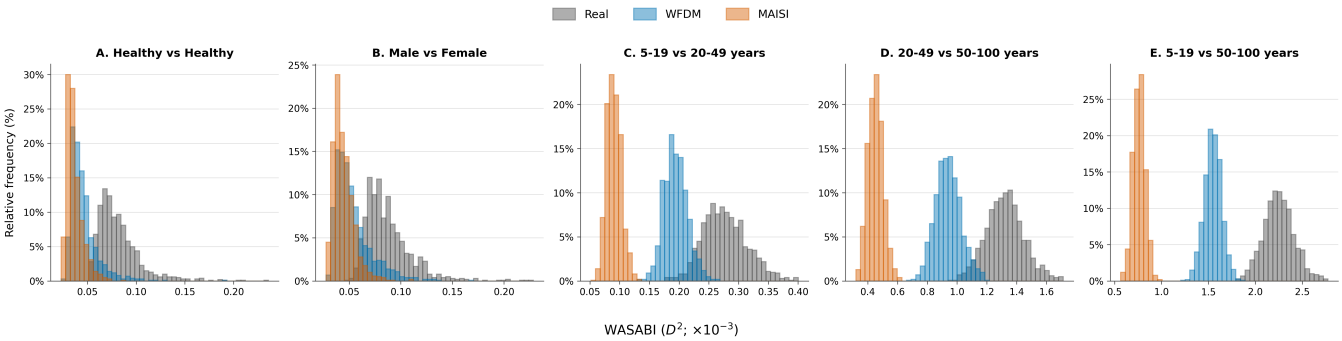


Figure 5: WASABI distributions for demographic morphometric comparisons in Real, WFDM, and MAISI T1w MRI. Panel A shows the same-population reference, comparing two randomly sampled groups within Real, WFDM, and MAISI separately. Panel B shows the male–female comparison separately for Real, WFDM, and MAISI. Panels C–E compare ages 5–19 versus 20–49, ages 20–49 versus 50–100, and ages 5–19 versus 50–100 years, respectively. The x-axis shows WASABI distance in units of 10^{-3}, and the y-axis shows relative frequency. Axis ranges are adjusted separately for each panel to show the distributions clearly.

4.6.3 Regional morphometric differences

Detailed eTIV-normalized regional volume distributions and signed Cohen's d values for all 52 measures are provided in Supplementary Material Appendix E. Across the 52 regional measures, WFDM showed a smaller $|d_{\text{Cohen}}|$ than MAISI in $45/52$ measures, including $32/34$ cortical measures and $13/18$ remaining global, subcortical, and ventricular volume measures. The largest WFDM cortical differences were observed in the posterior cingulate ($d_{\text{Cohen}} = -0.56$) and caudal anterior cingulate ($d_{\text{Cohen}} = -0.52$). In contrast, MAISI showed lower mean eTIV-normalized volume than Real in $33/34$ cortical regions, with the largest differences in rostral middle frontal ($d_{\text{Cohen}} = -1.55$), superior temporal ($d_{\text{Cohen}} = -1.44$), and superior frontal ($d_{\text{Cohen}} = -1.33$). Entorhinal and paracentral cortex were the only cortical regions in which MAISI showed a smaller $|d_{\text{Cohen}}|$ than WFDM.

Among the remaining 18 global, subcortical, and ventricular volume measures, the largest WFDM differences were observed in the inferior lateral ventricle ($d_{\text{Cohen}} = +0.66$), cerebellar white matter ($d_{\text{Cohen}} = +0.58$), and accumbens area ($d_{\text{Cohen}} = +0.54$). MAISI showed its largest differences in cerebral white matter ($d_{\text{Cohen}} = +1.85$), cerebral cortex ($d_{\text{Cohen}} = -1.70$), and cerebellar white matter ($d_{\text{Cohen}} = +1.52$). The five measures in which MAISI showed a smaller $|d_{\text{Cohen}}|$ were the inferior lateral ventricle, third ventricle, pallidum, amygdala, and total cerebrospinal fluid. Figure 6 shows the regional Cohen's d magnitudes, $|d_{\text{Cohen}}|$, mapped onto the corresponding anatomical regions. WFDM showed lower $|d_{\text{Cohen}}|$ across most mapped regions, whereas MAISI showed larger differences across multiple cortical and subcortical regions. When grouped by effect-size magnitude, $32/52$ (61.5%) WFDM regional measures showed small effects ($|d_{\text{Cohen}}| < 0.2$), $14/52$ (26.9%) showed small-to-medium effects

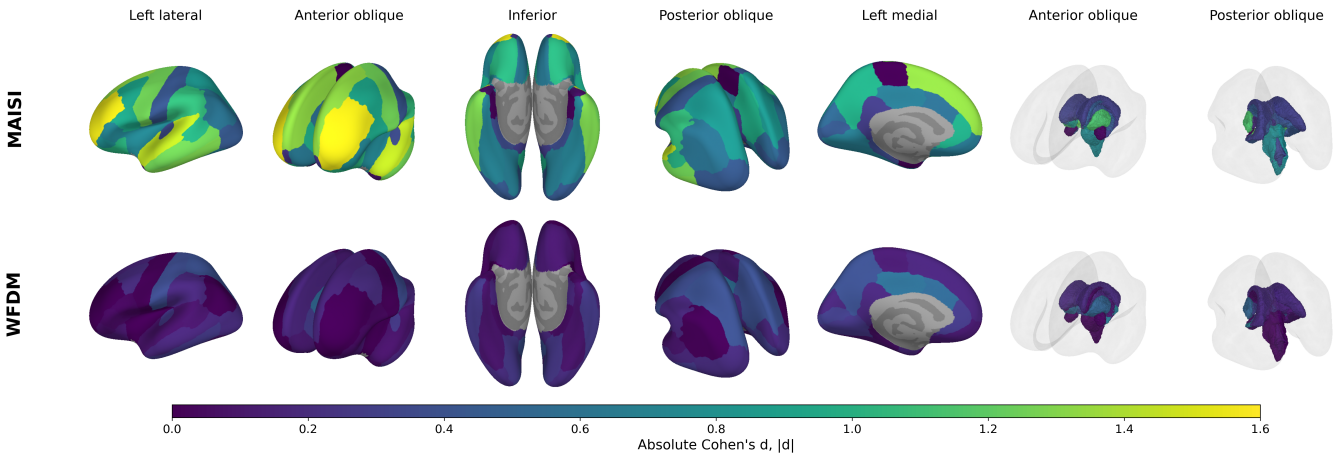


Figure 6: Regional Cohen's d magnitudes, $|d_{\text{Cohen}}|$, mapped onto the corresponding anatomical regions. The top row shows MAISI and the bottom row WFDM, with five cortical views followed by two oblique views of subcortical, ventricular, and brainstem structures. Colors represent $|d_{\text{Cohen}}|$ on a common 0–1.6 scale for both models, with lower values indicating smaller standardized differences from Real.

($0.2 \leq |d_{\text{Cohen}}| < 0.5$), and $6/52$ (11.5%) showed medium-to-large effects ($|d_{\text{Cohen}}| \geq 0.5$). For MAISI, the corresponding counts were $6/52$ (11.5%), $15/52$ (28.8%), and $31/52$ (59.6%), respectively (Peng et al., 2024; Sullivan & Feinn, 2012).

4.7 Metadata-associated regional morphometry

4.7.1 Age-associated regional morphometry

Figure 7 compares the 52 regional age–volume correlations in Real MRI with the corresponding correlations in WFDM and MAISI. WFDM showed stronger agreement with the Real regional age-effect pattern ($r = 0.982$, $p < 0.001$) than MAISI ($r = 0.866$, $p < 0.001$). The corresponding OLS fits had a slope of 1.143 and intercept of 0.030 for WFDM, compared with a slope of 0.866 and intercept of 0.125 for MAISI. WFDM preserved the direction of the Real age association in $51/52$ regions (98.1%), compared with $41/52$ (78.8%) for MAISI. The mean absolute regional correlation difference from Real was 0.051 for WFDM and 0.180 for MAISI. The only WFDM sign difference occurred in the frontal pole, where the Real association was close to zero (Real $r = 0.012$; WFDM $r = -0.016$). Among the strongest Real associations, cerebrospinal fluid volume increased with age (Real $r = 0.712$), which was preserved by WFDM ($r = 0.812$) and MAISI ($r = 0.681$). Cerebral cortex volume decreased with age (Real $r = -0.626$), with WFDM retaining a similar association ($r = -0.725$) and MAISI showing a weaker association ($r = -0.262$). The complete regional age–volume trajectories for Real, WFDM, and MAISI are provided in Supplementary Material Appendix F.

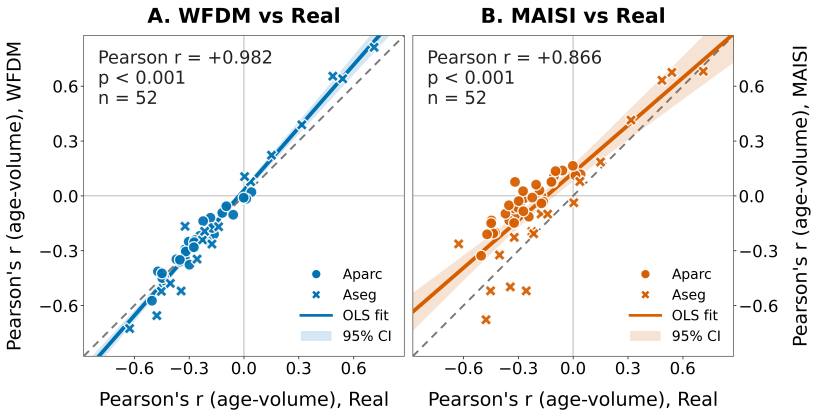


Figure 7: Preservation of age-associated regional morphometric patterns in synthetic T1w MRI. Panel A compares WFDM with Real MRI, and Panel B compares MAISI with Real MRI. Each point represents one of the 52 eTIV-normalized SynthSeg measures; circles denote the 34 bilateral cortical measures and crosses the remaining 18 measures. The x-axis shows the within-region Pearson correlation between age and regional volume in Real MRI, and the y-axis shows the corresponding correlation in the synthetic source. Positive correlations indicate increasing regional volume with age and negative correlations indicate decreasing volume. The dashed line shows identity, the solid line shows the OLS fit, and the shaded band shows its pointwise 95% confidence interval. The Pearson correlation coefficient in each panel quantifies the correspondence between Real and synthetic age–volume correlations across the 52 regional measures.

4.7.2 Sex-associated regional morphometry

Figure 8 compares the 52 regional male–female Cohen's d values in Real MRI with the corresponding values in WFDM and MAISI. WFDM showed stronger agreement with the Real regional effect-size pattern ($r = 0.922$, $p < 0.001$) than MAISI ($r = 0.865$, $p < 0.001$). The OLS fits had a slope of 0.851 and intercept of 0.022 for WFDM, compared with a slope of 1.036 and intercept of -0.027 for MAISI. WFDM preserved the direction of the Real male–female difference in $48/52$ regions (92.3%), compared with $44/52$ (84.6%) for MAISI. The mean absolute regional Cohen's d difference from Real was 0.092 for WFDM and 0.130 for MAISI. Among the strongest Real effects, frontal pole volume was larger in females (Real $d = -0.82$), with the direction and magnitude retained by WFDM ($d = -0.83$) and MAISI ($d = -1.05$). The middle temporal cortex showed the strongest positive Real effect (Real $d = 0.36$), which remained positive in WFDM ($d = 0.25$) and MAISI ($d = 0.14$). The complete regional sex-specific volume distributions and effect sizes are provided in Supplementary Material Appendix G.

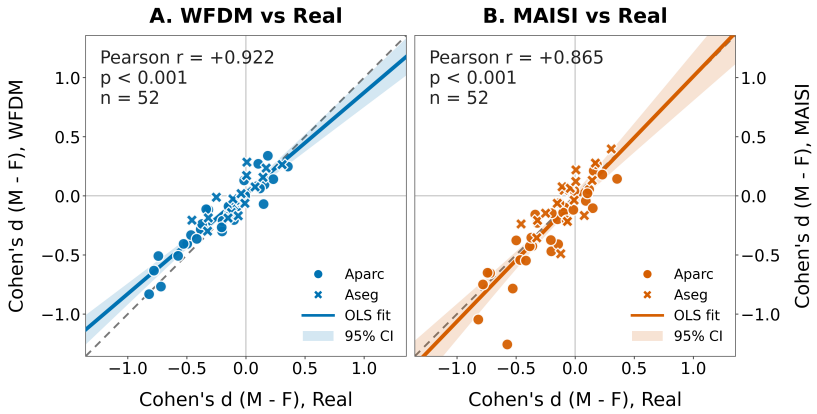


Figure 8: Preservation of sex-associated regional morphometric patterns in synthetic T1w MRI. Panel A compares WFDM with Real MRI, and Panel B compares MAISI with Real MRI. Each point represents one of the 52 eTIV-normalized SynthSeg measures; circles denote the 34 bilateral cortical measures and crosses the remaining 18 measures. The x-axis shows Cohen's d for Real MRI, and the y-axis shows the corresponding value in the synthetic source; effects are defined as male minus female. Positive values indicate greater regional volume in males, and negative values indicate greater volume in females. The dashed line shows identity, the solid line shows the OLS fit, and the shaded band shows its pointwise 95% confidence interval. The Pearson correlation coefficient in each panel quantifies the correspondence between Real and synthetic Cohen's d values across the 52 regional measures.

5 Discussion

The results of our study show that WFDM can synthesize T1w, T2w, T1Gd, and FLAIR MRI from heterogeneous MRI data using a single conditional 3D latent diffusion model. Within the matched BrainScape-trained comparison, WFDM achieved lower pooled FID and MMD than the MAISI internal baseline while maintaining comparable sample diversity. The modality-specific results indicate that target-modality conditioning enables a single generator to produce the requested MRI contrast, thereby supporting controllable multimodal synthesis in a heterogeneous pooled neuroimaging setting. WFDM addresses modality imbalance at the cohort level rather than performing subject-specific missing-modality completion. This distinction is relevant to BrainScape, where T1w MRI is widely represented while T2w, FLAIR, and particularly T1Gd have less consistent coverage across contributing datasets (Yasinzai et al., 2025). In its present form, WFDM generates a new MRI conditioned on the requested target modality and available metadata without requiring an additional MRI scan from the same subject as input. Subject-specific modality completion would additionally require conditioning on available MRI scans from the same subject, which could be implemented via an additional image-conditioning pathway such as a ControlNet adapter (Zhang et al., 2023).

The compressor results suggest that latent representation quality is an important factor in downstream synthesis performance. Among the learned compressors trained on BrainScape, WF-SA-VAE achieved

the best pooled performance across all reported reconstruction metrics, outperforming the BrainScape-trained MAISI VAE-GAN and the other BrainScape-trained candidates (Table 3). Although the released MAISI VAE-GAN achieved the strongest reconstruction performance among the learned compressors, the downstream synthesis results indicate that the compact WF-SA-VAE latent provides an effective representation for multimodal MRI synthesis on BrainScape. Under the same one-quarter spatial compression and four-channel latent, both WF-VAE and WF-SA-VAE substantially outperformed the Wavelet VAE across all reconstruction metrics, supporting the overall Wavelet Fusion Architecture under the evaluated configurations. Because the diffusion model operates on the compressed latent representation, anatomical detail or contrast information that is not preserved by the compression stage can limit downstream synthesis fidelity, reflecting the trade-off between complexity reduction and detail preservation in latent diffusion (Pinaya et al., 2022; Rombach et al., 2022). The stronger reconstruction fidelity of WF-SA-VAE relative to the BrainScape-trained compressors may therefore have contributed to WFDM's lower pooled FID and MMD in the matched generator comparison. A similar association is observed at the modality level, where T1Gd had the weakest WF-SA-VAE reconstruction fidelity and was also the most challenging WFDM synthesis modality, with the MAISI internal baseline achieving lower T1Gd FID and MMD.

The wavelet-fusion architecture ablation supports retaining a second wavelet stage, as all two-stage configurations improved LPIPS and SSIM relative to the Stage-1 only configuration (Table 8). Although Stage-2 LLL outperformed Low4 on most reconstruction metrics, we retained Stage-2 Low4 based on the architectural rationale of incorporating wavelet information from four selected subbands at the deeper encoder stage, while avoiding the larger Stage-2 coefficient representation produced by decomposing all eight subbands. The Stage-2 wavelet decomposition is deterministic and requires no training, while Stage-2 Low4 also achieved the lowest LPIPS and maintained SSIM and MS-SSIM close to the best-performing configurations. The observed ranking may also depend on the Stage-2 projection capacity, which was not varied in this ablation. The sampling comparison in Experiment C showed that the 50-step RFlow configuration achieved lower pooled FID and MMD than the evaluated 1000-step DDPM variants while substantially reducing inference time. A similar result was reported for CT synthesis in MAISI-v2, where 30-step RFlow inference achieved a slightly lower average FID than the 1000-step MAISI DDPM baseline while substantially reducing latent-diffusion inference time (Zhao et al., 2025). The MAISI-v2 study also explored the effect of RFlow step count and found that average FID improved as the number of inference steps increased from 5 to 100, while the additional gains became progressively smaller as more steps were used. The authors therefore selected 30 steps as a practical trade-off between synthesis quality and inference cost (Zhao et al., 2025).

WFDM also includes demographic and clinical metadata conditioning with an explicit metadata availability mask. The raw metadata are first transformed into a fixed-length numerical vector and corresponding

availability mask, which are then processed by a learnable metadata encoder to produce the conditioning representation used by the diffusion model. This design is important for BrainScape because metadata coverage varies across contributing datasets (Yasinzai et al., 2025), and the availability mask allows observed covariates to be distinguished from unavailable fields without requiring every source dataset to provide the same demographic and clinical variables. This is particularly important for clinical indicators, which are available only for subsets of the contributing datasets and therefore represent available source-dataset annotations rather than complete cohort-wide labels across BrainScape.

Synthetic MRI evaluation remains difficult to standardize because reported metric values can vary with both metric choice and evaluation implementation (Dohmen et al., 2025; Koetzier et al., 2024). To provide a consistent evaluation protocol for 3D MRI synthesis, we use and publicly release *medmetric*, with FID and MMD computed in a fixed 3D MedicalNet feature space and MS-SSIM reported separately in image space as a diversity proxy (Yasinzai, 2026b). This common evaluation pipeline provides a consistent basis for the within-study comparisons, while comparisons with studies using different feature extractors, normalization procedures, or slice-wise versus 3D metric computation should be interpreted with caution (Dohmen et al., 2025; Koetzier et al., 2024).

The morphometric analyses show that WFDM more closely reproduces the morphometric characteristics of Real MRI than the BrainScape-trained MAISI baseline. WFDM showed lower WASABI distance than MAISI across T1w, T2w, and FLAIR, with the T1w distance remaining close to the within-real reference. Regional T1w analysis was consistent with this finding, with WFDM showing smaller standardized regional differences relative to Real than MAISI across most evaluated regions. WFDM also more closely preserved age-associated morphometric separation, the regional age–volume association pattern, and the regional male–female effect pattern observed in Real MRI. These analyses support population-level preservation of age- and sex-associated morphometric structure. However, randomly sampled groups from both synthetic T1w cohorts showed smaller WASABI distances than groups drawn from Real MRI, suggesting reduced population-level morphometric variation. This interpretation is also consistent with the reduced age-group separations observed for WFDM, which reached approximately 70% of the corresponding Real values.

Although WFDM is trained on MRI from publicly available BrainScape source datasets (Yasinzai et al., 2025), synthetic outputs should not be assumed to provide an inherent privacy guarantee. Generative models can memorize training examples or leak information about the training data distribution, so synthetic MRI should not be assumed safe for unrestricted sharing without appropriate privacy evaluation (Daum et al., 2024; Shi et al., 2025; Usman Akbar et al., 2025). Another concern is the recursive use of synthetic data to train subsequent generative models. Repeated training on synthetic data can

contribute to model collapse, with errors accumulating across generations and low-probability regions of the original distribution progressively disappearing (Shumailov et al., 2024). In medical imaging, this loss is particularly concerning for rare diseases, underrepresented demographic subgroups, uncommon acquisition protocols, and underrepresented MRI contrasts, which may already be sparsely represented in the training data. Accordingly, synthetic MRI should retain clear provenance and be used as a supplement to real data, with future generative training anchored to real MRI rather than relying recursively on synthetic outputs.

5.1 Limitations

A first limitation is that externally pretrained synthesis models do not constitute a controlled architectural comparison. Only WFDM and the MAISI internal baseline were trained on the same BrainScape split using the same target-modality and metadata-conditioning interface. LDM, WDM, HA-GAN, and Med-DDPM retain their released training datasets, conditioning mechanisms, objectives, and output formulations and are therefore included as contextual external references. Their native formulations also differ: the published LDM generates T1w MRI using age, sex, ventricular volume, and normalized brain volume as conditioning inputs (Pinaya et al., 2022), WDM and the primary HA-GAN brain-synthesis configurations are unconditional T1w generators (Friedrich et al., 2024; Sun et al., 2022), and Med-DDPM uses segmentation-mask conditioning to jointly generate four BraTS modalities (Dorjsembe et al., 2024). Matching these models to the BrainScape target-modality, metadata, and metadata-availability conditioning interface would therefore require method-specific conditioning modifications and, for some architectures, changes to the output formulation. Single-modality generators would either need to be modified to support explicit conditioning on the target modality or trained as separate models for each target contrast. Both strategies would require complete retraining, method-specific tuning, substantial implementation effort, considerable computational resources, and additional training time. The performance of these adapted variants would reflect both the original architectures and the specific adaptation and optimization choices. A fully standardized benchmark of such adapted synthesis architectures would therefore be valuable, but is beyond the scope of the present study. The same limitation applies to the compressor benchmark: externally pretrained compressors are included as contextual references, while compressor selection is based on the candidates trained on BrainScape.

A second limitation is that the metadata-associated morphometric analyses assess preservation of observed population-level associations and do not isolate metadata controllability. Although WFDM preserved age- and sex-associated regional patterns, the study does not test the effect of varying a single metadata field while holding the other conditioning inputs fixed. A third limitation is that WFDM is

a cohort-level generator rather than a subject-specific modality-completion model. Although target-modality conditioning specifies the contrast to be generated, WFDM does not condition on available MRI scans from the same subject. Subject-specific missing-contrast synthesis would require an additional image-conditioning pathway to provide anatomical guidance from those scans (Zhang et al., 2023). Finally, T1Gd remains the most challenging target modality for WFDM. Its weaker synthesis performance may be related to both the substantially smaller T1Gd representation in the BrainScape training data and the comparatively weaker T1Gd reconstruction fidelity of WF-SA-VAE. Although weighted sampling is used to mitigate target-modality imbalance during training, it does not increase the number or diversity of available T1Gd scans. Future work should therefore examine whether expanding the T1Gd training set and improving the learned latent representation can enhance T1Gd synthesis performance.

5.2 Future work

A primary future direction is to extend WFDM from controllable multimodal MRI synthesis to subject-specific modality completion. This could be achieved by adding an image-conditioning pathway, such as a ControlNet adapter, that uses available MRI scans from the same subject to guide synthesis of the missing target contrast (Zhang et al., 2023). The existing target-modality and metadata-conditioning pathways would continue to specify the requested contrast and provide available metadata context, while image conditioning would add subject-specific anatomical guidance. This extension could support the creation of an augmented multimodal BrainScape dataset for downstream tasks that require complete modality sets, such as brain tumor segmentation.

Future work should also evaluate metadata controllability by examining whether varying demographic and clinical metadata produces consistent and interpretable changes in generated MRI. A further direction is to extend the synthesis benchmark by adapting and training additional architectures on a common BrainScape training split using equivalent target-modality and metadata information, with evaluation under a common protocol. Any model-specific adaptations required for this benchmark should be clearly documented.

6 Data availability

This study uses BrainScape, an open-source framework that aggregates public brain MRI datasets through dataset-specific configuration files and metadata (Yasinzai et al., 2025). Researchers should refer to the BrainScape paper for details on data access, dataset reconstruction, and the preprocessing workflow used

to generate the study data from the original public sources.

7 Code availability

The code and supporting resources for this study are available through the project repository at https://github.com/yasinzaii/generative-mri. The repository serves as a public gateway to the study source code, model configurations, training and sampling workflows, environment specifications, and accompanying software packages.

Reusable model, data-handling, conditioning, inference, and evaluation components developed for this study are released through *BrainMint*, an open-source Python package available from GitHub at https://github.com/yasinzaii/brainmint (Yasinzai, 2026a). The evaluation workflows use *medmetric*, which provides the standardized metric implementations used for synthetic MRI evaluation in this study (Yasinzai, 2026b).

Author Contributions

1. **Muhammad Nabi Yasinzai (MNY)**: Conceptualization, Methodology, Software, Writing – Original Draft.
2. **Remika Mito (RM)**: Conceptualization, Supervision, Writing – Review & Editing.
3. **Mangor Pedersen (MP)**: Conceptualization, Supervision, Writing – Review & Editing, Resources, Funding Acquisition.

All authors read, revised, and approved the final version of the manuscript.

Funding

MNY is a recipient of the Australian Epilepsy Project PhD scholarship (Medical Research Future Fund - Frontier Health and Medical Research Program - Grant Number MRFF75908). RM is the recipient of an Australian Research Council Discovery Early Career Researcher Award (project number: DE240101035). MP is a recipient of Health Research Council fellowship, New Zealand (#21/622).

Declaration of Competing Interests

The authors declare that they have no competing financial or non-financial interests that could have influenced the work reported in this manuscript.

Generative AI statement

During the preparation of this manuscript, the authors used ChatGPT to improve grammar, sentence structure, clarity, and readability. The authors reviewed and edited all AI-assisted text and take full responsibility for the content of the publication.

Acknowledgements

We thank the participants in the original studies whose data make this work possible. Their contributions to open neuroimaging resources form the foundation of BrainScape's mission to build a broad, diverse, and evolving dataset.

We also acknowledge New Zealand eScience Infrastructure (NeSI) for providing the high-performance computing resources that supported the training and evaluation of the models in this study.

References

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R. T., Berger, C., Ha, S. M., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., ... Albiol. (2018). Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*.

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

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Coupé, P., Hellier, P., Prima, S., Kervrann, C., & Barillot, C. (2008). 3d wavelet subbands mixing for image denoising. *International Journal of Biomedical Imaging*, *2008*(1), 590183.

Dadar, M., & Duchesne, S. (2020). Reliability assessment of tissue classification algorithms for multi-center and multi-scanner data. *NeuroImage*, *217*, 116928.

Daum, D., Osuala, R., Riess, A., Kaissis, G., Schnabel, J. A., & Di Folco, M. (2024). On differentially private 3d medical image synthesis with controllable latent diffusion models. *MICCAI Workshop on Deep Generative Models*, 139–149.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, *31*(3), 968–980.

Dohmen, M., Klemens, M. A., Baltruschat, I. M., Truong, T., & Lenga, M. (2025). Similarity and quality metrics for mr image-to-image translation. *Scientific Reports*, *15*(1), 3853.

Dorjsembe, Z., Pao, H.-K., Odonchimed, S., & Xiao, F. (2024). Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, *28*(7), 4084–4093.

Fan, Y., Liao, H., Huang, S., Luo, Y., Fu, H., & Qi, H. (2024). A survey of emerging applications of diffusion probabilistic models in mri. *Meta-Radiology*, *2*(2), 100082.

Fischl, B. (2012). Freesurfer. *Neuroimage*, *62*(2), 774–781.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355.

Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., & Cattin, P. C. (2024). Wdm: 3d wavelet diffusion models for high-resolution medical image synthesis. *MICCAI workshop on deep generative models*, 11–21.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. S. (2013). Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, *7*, 267.

Guo, P., Zhao, C., Yang, D., Xu, Z., Nath, V., Tang, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., & Xu, D. (2025). Maisi: Medical ai for synthetic imaging. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4430–4441.

Huisman, S., Maspero, M., Philippens, M., Verhoeff, J., & David, S. (2024). Validation of synthseg segmentation performance on ct using paired mri from radiotherapy patients. *NeuroImage*, *303*, 120922.

Jafrasteh, B., Peng, W., Wan, C., Luo, Y., Adeli, E., & Zhao, Q. (2025). Wasabi: A metric for evaluating morphometric plausibility of synthetic brain mris. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 684–694.

Jiang, L., Mao, Y., Wang, X., Chen, X., & Li, C. (2023). Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 398–408.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *European conference on computer vision*, 694–711.

Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., & Merhof, D. (2023). Diffusion models in medical imaging: A comprehensive survey. *Medical image analysis*, *88*, 102846.

Kim, J., & Park, H. (2024). Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. *Proceedings of the IEEE/CVF Winter conference on applications of computer Vision*, 7604–7613.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koetzier, L. R., Wu, J., Mastrodicasa, D., Lutz, A., Chung, M., Koszek, W. A., Pratap, J., Chaudhari, A. S., Rajpurkar, P., Lungren, M. P., & Willemink, M. J. (2024). Generating synthetic data for medical imaging. *Radiology*, *312*(3), e232471.

Liu, P., Zhang, H., Zhang, K., Lin, L., & Zuo, W. (2018). Multi-level wavelet-cnn for image restoration. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 773–782.

Lüpke, S., Yeganeh, Y., Adeli, E., Navab, N., & Farshad, A. (2024). Physics-informed latent diffusion for multimodal brain mri synthesis. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 198–207.

Meng, X., Sun, K., Xu, J., He, X., & Shen, D. (2024). Multi-modal modality-masked diffusion network for brain mri synthesis with random modality missing. *IEEE Transactions on Medical Imaging*, *43*(7), 2587–2598.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., ... Leemput, V. (2014). The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, *34*(10), 1993–2024.

Miao, Y., Deng, J., & Han, J. (2024). Waveface: Authentic face restoration with efficient frequency recovery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6583–6592.

Peng, W., Bosschieter, T., Ouyang, J., Paul, R., Sullivan, E. V., Pfefferbaum, A., Adeli, E., Zhao, Q., & Pohl, K. M. (2024). Metadata-conditioned generative models to synthesize anatomically-plausible 3d brain mris. *Medical Image Analysis*, *98*, 103325.

Phung, H., Dao, Q., & Tran, A. (2023). Wavelet diffusion models are fast and scalable image generators. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10199–10208.

Pinaya, W. H., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., & Cardoso, M. J. (2022). Brain imaging generation with latent diffusion models. *MICCAI workshop on deep generative models*, 117–126.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shi, Y., Xia, W., Niu, C., Wiedeman, C., & Wang, G. (2025). Privacy-preserving latent diffusion-based synthetic medical image generation. *IEEE Transactions on Medical Imaging*.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). Ai models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755–759.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of graduate medical education*, *4*(3), 279–282.

Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., & Batmanghelich, K. (2022). Hierarchical amortized gan for 3d high resolution medical image synthesis. *IEEE journal of biomedical and health informatics*, *26*(8), 3966–3975.

Tudosiu, P.-D., Pinaya, W. H. L., Ferreira Da Costa, P., Dafflon, J., Patel, A., Borges, P., Fernandez, V., Graham, M. S., Gray, R. J., Nachev, P., Ourselin, S., & Cardoso, M. J. (2024). Realistic morphology-preserving generative modelling of the brain. *Nature Machine Intelligence*, *6*(7), 811–819.

Usman Akbar, M., Wang, W., & Eklund, A. (2025). Beware of diffusion models for synthesizing medical images—a comparison with gans in terms of memorizing brain mri and chest x-ray images. *Machine Learning: Science and Technology*, *6*(1), 015022.

Yasinzai, M. N. (2026a). *Brainmint: Brain mri synthesis, compression, modality translation, and evaluation toolkit* (Version 0.1.1) [Open-source Python package]. https://github.com/yasinzaii/brainmint

Yasinzai, M. N. (2026b). *Medmetric: Metrics for synthetic mri generation* (Version v0.1.1). https://github.com/yasinzaii/medmetric/releases/tag/v0.1.1

Yasinzai, M. N., Mito, R., & Pedersen, M. (2025). Brainscape: An open-source framework for integrating and preprocessing anatomical mri datasets. *Imaging Neuroscience*, *3*, IMAG–a.

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF international conference on computer vision*, 3836–3847.

Zhang, X., Liang, J., Cao, P., Yang, J., & Zaiane, O. R. (2025). Structure-aware mri translation: Multi-modal latent diffusion model with arbitrary missing modalities. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 508–518.

Zhao, C., Guo, P., Yang, D., Tang, Y., He, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., & Xu, D. (2025). Maisi-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. *arXiv preprint arXiv:2508.05772*.

Zhou, Z., & Lu, Z.-R. (2013). Gadolinium-based contrast agents for magnetic resonance cancer imaging. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*, *5*(1), 1–18.

Zhu, L., Xue, Z., Jin, Z., Liu, X., He, J., Liu, Z., & Yu, L. (2023). Make-a-volume: Leveraging latent diffusion models for cross-modality 3d brain mri synthesis. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 592–601.