

Diffusion MRI Transformer with a Diffusion Space Rotary Positional Embedding (D-RoPE)

Gustavo Chau Loo Kung Mohammad Abbasi Camila Blank Juze Zhang Alan Q. Wang
 Sophie Ostmeier Akshay Chaudhari Kilian Pohl Ehsan Adeli
 Stanford University

Abstract

Diffusion Magnetic Resonance Imaging (dMRI) plays a critical role in studying microstructural changes in the brain. It is, therefore, widely used in clinical practice; yet progress in learning general-purpose representations from dMRI has been limited. A key challenge is that existing deep learning approaches are not well-suited to capture the unique properties of diffusion signals. Brain dMRI is normally composed of several brain volumes, each with different attenuation characteristics dependent on the direction and strength of the diffusion-sensitized gradients. Thus, there is a need to jointly model spatial, diffusion-weighting, and directional dependencies in dMRI. Furthermore, varying acquisition protocols (e.g., differing numbers of directions) further limit traditional models. To address these gaps, we introduce a diffusion space rotatory positional embedding (D-RoPE) plugged into our dMRI transformer to capture both the spatial structure and directional characteristics of diffusion data, enabling robust and transferable representations across diverse acquisition settings and an arbitrary number of diffusion directions. After self-supervised masked autoencoding pretraining, tests on several downstream tasks show that the learned representations and the pretrained model can provide competitive or superior performance compared to several baselines in these downstream tasks (even compared to a fully trained baseline); the finetuned features from our pretrained encoder resulted in a 6% higher accuracy in classifying mild cognitive impairment and a 0.05 increase in the correlation coefficient when predicting cognitive scores. Code is available at: github.com/gustavochau/D-RoPE.

1. Introduction

Diffusion Magnetic Resonance Imaging (dMRI) is a non-invasive imaging technique that is routinely used to probe the diffusion of water molecules inside the brain and other organs [9], and is widely used in clinical practice [22]. dMRI provides information about white matter integrity

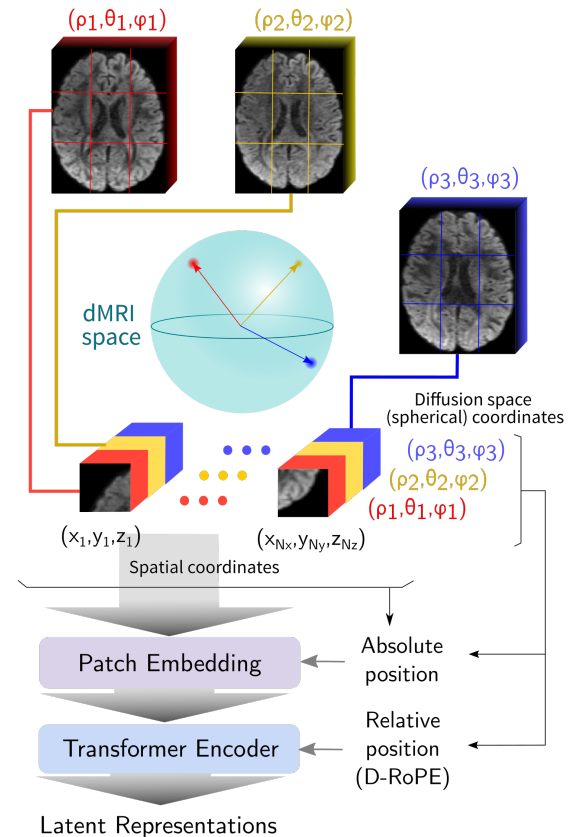


Figure 1. Overview of dMRI data structure and model input. The purple sphere represents diffusion sampling directions at a specific b-value. 3D dMRI volumes acquired at different directions are divided into 3D patches. Spatial coordinates (x, y, z) define the image domain and spherical coordinates (ρ, θ, φ) describe the diffusion space. Patches are linearized with absolute positional embeddings and processed by a transformer encoder incorporating relative positional encoding (D-RoPE; see Section 3.1 and Fig. 3).

[2, 36], putative structural connections [35, 48], and tissue microstructure [6, 51]. In addition, metrics derived from this information are able to identify the effects of certain brain diseases, such as Alzheimer’s [13, 53] and multiple sclerosis [24]. In dMRI, the molecular diffusion in a given

voxel is encoded in the attenuation of the MRI signal and is dependent on the microstructure, the direction of the diffusion encoding gradient (normally called b-vector), and the “strength” of the diffusion encoding (normally expressed as a b-value). A normal pulsed gradient spin echo (PGSE) dMRI acquisition [59] can be viewed as a series of brain volumes, each corresponding to a point on the surface of a sphere. The diameter of the sphere comes from the b-value, while the position of the point in the sphere corresponds to the b-vector (see Figure 1).

While there have been extensive advances in novel architectures and representation learning for other medical imaging and MRI modalities [15, 43, 54, 55], the progress in obtaining general-purpose representations from dMRI has been limited [32, 56], or the representations have only focused on specific applications [64, 68]. A key challenge is that existing deep learning approaches are not well-suited to capture the unique properties of the spatial information and directional dependencies in diffusion signals. Additionally, the problem is further complicated by the multiple and varying diffusion protocols (e.g., different numbers of diffusion encoding directions and b-values) often used to acquire dMRI. Previous attempts at dealing with these challenges for specific applications include: upsampling to a common protocol [27], modeling the diffusion space as a graph [16, 17, 69], using modified convolutional operations [44], or concatenating the protocol information as part of the input [72].

As a result, there is a need for architectures that can model the joint spatial and diffusion space dependencies of dMRI and that can handle a variety of dMRI protocols to obtain general and transferable representations. In this paper, we propose a model that combines novel absolute and relative positional embeddings to encode the geometric information of the different volumes (Figure 1). The proposed model is able to handle acquisitions with multiple b-values and an arbitrary number of directions.

In summary, the contributions of this work are as follows: (i) we propose a novel architecture inspired by video transformers that can obtain representations from arbitrary dMRI acquisitions; (ii) we propose D-RoPE, which augments the attention mechanism with a generalization of Rotary Positional Embedding (RoPE) that models the interdependence of different diffusion directions and b-values; and (iii) we perform a thorough evaluation of the architecture and test the representations in several downstream tasks (Age regression, sex classification, mild cognitive impairment (MCI) classification, and Alzheimer’s disease assessment scale (ADAS) regression).

2. Related Work

Task-specific models for diffusion MRI. Some previous work focused on obtaining high-resolution reconstructions

of the diffusion tensor [38], or of maps of diffusion tensor metrics (e.g. fractional anisotropy, mean diffusivity) or other microstructural parameters [71, 72]. Other works have focused on obtaining meaningful information from the dMRI images or dMRI-derived connectomes [17, 18, 69]. Though these and other studies have shown the effectiveness of deep learning for dMRI, there is a need to develop models that can be adapted for a wide variety of tasks.

Representation Learning for Brain MRIs. Foundation models have been proposed to extract representations of structural (T1 and T2 weighted) brain MRI [8, 61]. Another recent publication combined structural MRI and diffusion MRI for different tasks [37], including tractography streamline generation. Foundation models have also been proposed for other modalities, like functional MRI (fMRI) [15, 20]. For dMRI, existing representation learning work has focused on superresolution [64], microstructure reconstruction [68], and tractography features or representations [32, 56], but there has been a lack of work focusing on obtaining general and transferable representations from raw dMRI data.

Attention in Video Transformers There have been different strategies for adapting attention modules for video inputs, so as to simultaneously handle attention in the space and time domains. [11] showed one of the first attempts by alternating attention in space and time. Prior works looked at alternative ways of integrating both attentions with reduced complexity via space-time mixing [14], or via tubelets and factorized attention [5].

Rotary Positional Embedding [60] introduced the concept of Rotary Positional Embedding (RoPE) to integrate information about the relative position of tokens in the computation of the attention scores. Since then, some works have looked at different ways of adapting RoPE to the multi-dimensional case of video [63, 66]. Recently, LieRE [52] was proposed to generalize RoPE to high-dimensional rotation matrices for 2D and 3D vision tasks. Despite all this progress, none of this relative positional information can be applied directly to dMRI data, which is normally represented in a spherical space.

3. Methods

We adopt a transformer model architecture and train it with the Masked Autoencoding (MAE) [30] objective as summarized in Figure 2. Given a series of N_d diffusion volumes of size (N_x, N_y, N_z) , the input is divided into 3D patches of size (P_x, P_y, P_z) , as shown in Figure 1, and then linearized to tokens of latent dimension d . Thus, the resulting patch embedding is of size $S \times N_d \times d$, where $S = \frac{N_x}{P_x} \times \frac{N_y}{P_y} \times \frac{N_z}{P_z}$.

Vision Transformer (ViT) [21] architectures typically use a learnable positional encoding [38, 71] or a fixed sinusoidal positional encoding [21]; however, this approach will be inadequate for samples with different dMRI protocol in-

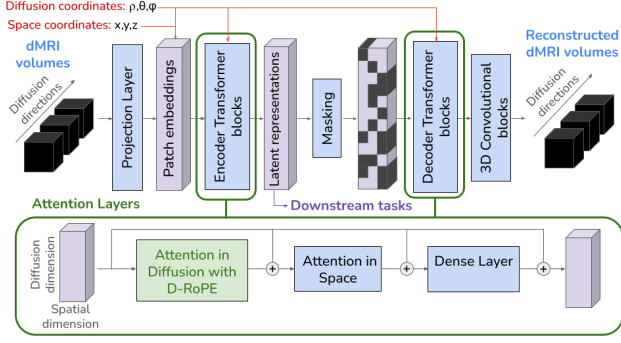


Figure 2. Summary of the model architecture. The different diffusion volumes are patchified via a projection layer and the absolute positional information of the image space and the diffusion space is added to the patch embeddings. The tokens are passed through modified attention blocks that alternate attention in both image and diffusion spaces. Additionally, relative information of different diffusion directions and b-values is encoded via D-RoPE. The obtained latent representations from the encoder are used for downstream task evaluation. To close the pretraining loop, some tokens are masked and fed into transformer blocks as the ones used in the encoder, followed by 3D convolutional blocks to reconstruct the original volumes.

formation. In our model, the absolute positional encoding is modified to include both spatial and diffusion-space information. The spatial position (x, y, z) is encoded using a 3D sinusoidal positional encoding into a $d/2$ dimensional vector p_s . The diffusion direction information, expressed in spherical coordinates (ρ, θ, φ) , is encoded using a learnable linear layer into another $d/2$ dimensional vector p_d . Both vectors, p_s and p_d , are concatenated and then summed to the patch embeddings. The embeddings are fed to the transformer encoder, which consists of a modified ViT [21]. Inspired by video transformers [11], the attention blocks alternate between the spatial and diffusion dimensions. The attention blocks also include a modified relative positional encoding (details in Section 3.1), which enables the model to exploit the interdependence between the different diffusion volumes. With this structure of the positional embedding and the attention blocks, the encoder supports an arbitrary number of diffusion volumes.

The encoder output provides the latent representation used for downstream tasks. To close the pretraining loop, some tokens are masked and fed into transformer blocks as the ones used in the encoder, followed by 3D convolutional blocks to reconstruct the original volumes. The loss function is a combination of the Mean Squared Error (MSE) of the masked and unmasked patches:

$$L = (1 - \tau)L_{\text{unmasked}} + \tau L_{\text{masked}}, \quad (1)$$

where τ is linearly ramped from 0.05 to 0.95 across epochs.

3.1. D-RoPE

The original Rotary Positional Embedding (RoPE) [60] modifies the attention mechanism by applying a rotation matrix $R_{\alpha, m-n}^d$ to encode the exact relative distance $m-n$ between the m th query, q_m , and the n -th key, k_n :

$$\mathbf{R}_{\alpha, \Delta}^d = \begin{pmatrix} \cos \Delta \alpha_1 & -\sin \Delta \alpha_1 & \cdots & 0 & 0 \\ \sin \Delta \alpha_1 & \cos \Delta \alpha_1 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \cos \Delta \alpha_{d/2} & -\sin \Delta \alpha_{d/2} \\ 0 & 0 & \cdots & \sin \Delta \alpha_{d/2} & \cos \Delta \alpha_{d/2} \end{pmatrix}, \quad (2)$$

where $\Delta = m - n$, d is the number of embedding dimensions, and $\alpha_i = 10000^{-2(i-1)/d}$, $i \in [1, 2, \dots, d/2]$. The similarity between q_m and k_n involves the inner product:

$$\langle q_m, k_n \rangle = q_m^\top \mathbf{R}_{\alpha, m-n}^d k_n. \quad (3)$$

Due to the block diagonal structure of $\mathbf{R}_{\alpha, m-n}^d$ with 2×2 rotations blocks, efficient implementations apply element-wise rotations directly to the query vector x_m and key vectors x_n .

We can think of $\mathbf{R}_{\alpha, m-n}^d$ as encoding relative distances, $(m-n)$ of a 1D sequence. However, dMRI volumes have a more complex spatial structure: each 3D volume is associated with a point $(b, \mathbf{v}) \in \mathbb{R}^+ \times \mathbb{S}^2$ in the diffusion space, where b is the b-value and $\mathbf{v}$ is the diffusion gradient direction (b-vector), and such the linear and ordinal distance used in RoPE is not appropriate for this case. To generalize RoPE, we substitute $(m-n)$ by a specifically designed distance $D(m, n)$. As a result, (2) becomes (the m and n dependency is omitted for clarity):

$$\mathbf{R}_{\theta, m, n}^d = \begin{pmatrix} \cos D\alpha_1 & -\sin D\alpha_1 & \cdots & 0 & 0 \\ \sin D\alpha_1 & \cos D\alpha_1 & \cdots & 0 & 0 \\ 0 & 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \cos D\alpha_{d/2} & -\sin D\alpha_{d/2} \\ 0 & 0 & \cdots & \sin D\alpha_{d/2} & \cos D\alpha_{d/2} \end{pmatrix}. \quad (4)$$

The distance $D(m, n)$ models the geometry of the diffusion space. We consider that two tokens x_m and x_n correspond to two diffusion acquisitions represented by $(b_n, \mathbf{v}_n), (b_m, \mathbf{v}_m) \in \mathbb{R}^+ \times \mathbb{S}^2$. That is, the corresponding b-values are defined in a linear space and the b-vectors are defined on the unitary sphere surface. Then, we define $D(m, n)$ as:

$$D(m, n) = D((b_m, \mathbf{v}_m), (b_n, \mathbf{v}_n)) \quad (5)$$

$$= \sqrt{\gamma(b_m - b_n)^2 + \arccos^2(|\mathbf{v}_m \cdot \mathbf{v}_n|)}, \quad (6)$$

where γ is a non-negative parameter that allows for weighting the contributions of the different components. Note that the absolute value inside the arc cosine provides 180° symmetry, where opposite gradient directions ($\mathbf{v}$ and $-\mathbf{v}$) are treated as identical, since they produce equivalent diffusion encoding. The concept of this diffusion space rotary positional embedding (D-RoPE), is summarized in Figure 3.

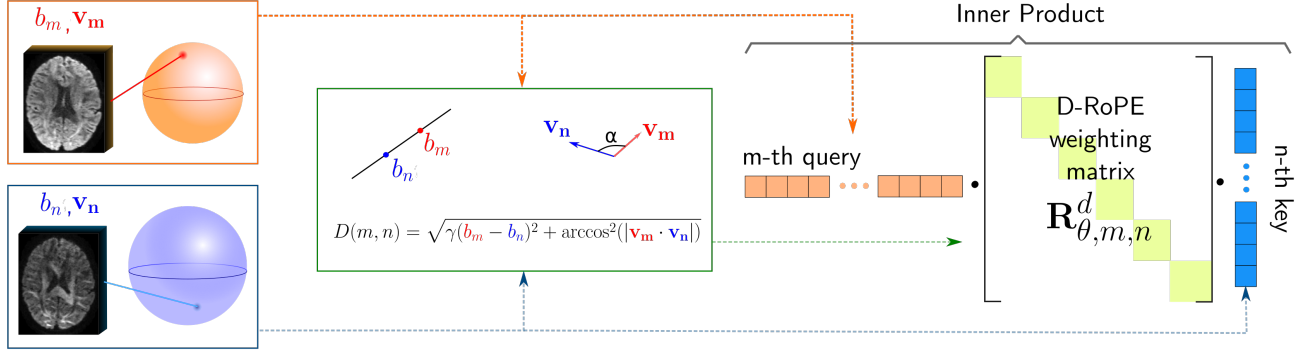


Figure 3. Attention calculation with D-RoPE: The relative rotation matrix depends on the distance between diffusion volumes, which is defined as a weighted combination of the distance between b-values and the angle between b-vectors.

3.2. Datasets and preprocessing

To train and evaluate our model, we used the combination of datasets from the Human Connectome Project (HCP) [23, 58] and the human Alzheimer’s Disease Neuroimaging Initiative (ADNI) [49]. First, the HCP Young Adult (HCP-YA) [23] dataset (n=1,065) was used for the pretraining stage and evaluation of the MAE reconstruction task with a 70/15/15 split for training/validation/test. For downstream evaluation, HCP Development (HCP-D, n=612) [58], HCP Aging (HCP-A, n=708) [12], and the fourth phase of ADNI (n=276) [49] were used under a cross-validation setting. Experiments are detailed in Section 4.3.

For HCP-A, HCP-D, and ADNI, distortion correction was conducted using Topup [4, 57] followed by Eddy current correction [3] using our publicly available preprocessing pipeline [1] and FSL [33]. HCP-YA data already include these two steps [28]. In case of multiple scans per subject, the different scans were rigidly registered and merged. Finally, the merged dMRI volumes were registered by a combination of rigid and affine transformations to MNI space using ANTs [7] to account for possible confounders of brain size and volume. To be input to the network, the diffusion weighted images were normalized by the b=0 reference volume, and outliers due to numerical errors were clipped (No logarithmic transformation was applied).

4. Experiments

4.1. Masked AutoEncoder Pretraining

We considered three different masking strategies: (i) normal spatial masking ratio of 75%, (ii) masking in the diffusion direction (i.e., random directions are entirely masked) with a ratio of 50%, and (iii) an alternation of the previous two on an epoch-by-epoch basis. In the following subsections, they are referred to as spatial, diffusion, and alternating, respectively. Ten encoding layers and three decoding layers were used, followed by three 3D convolutional lay-

ers, with an embedding dimension size of 384. The MAE was trained with a learning rate and weight decay following a cosine schedule. A patch size of (8,8,4) was used for all experiments (An evaluation of other patch sizes is shown in section 6.2 in the supplementary material). At each epoch, slabs of four contiguous slices and 15 random directions are selected for each patient to reduce memory consumption and act as a form of data augmentation. A detailed table with the hyperparameters used during the pretraining stage is shown in Table 5 in the supplementary material.

4.2. Ablation and Evaluation

In addition to the three different versions of masking strategies (spatial, diffusion, and alternating), we evaluated both the MAE with and without D-RoPE. We compared reconstructed volumes against the ground-truth volumes in the test set after a 75% spatial masking in 90 randomly selected directions. The reconstructions were compared in terms of Peak signal-to-noise-ratio (PSNR), Structural Similarity Image Metric (SSIM), and Fréchet Inception distance (FID) [31]. For the FID, the features of the different layers of a pretrained 2D ResNet50 [29] architecture were used. Additionally, we fitted the diffusion tensor model to the reconstructed volumes of each method and obtained fractional anisotropy (FA) and mean diffusivity (MD) maps using DIPY [26]. We compared the mean absolute error between these fitted maps and the ground-truth maps in several known white matter tracts (e.g., corpus callosum, inferior longitudinal fasciculus, etc.) to test the anatomical plausibility of the generated images. The segmentation of the tracts was obtained using TractSeg [65] on the ground-truth data.

4.3. Downstream Evaluations

We tested the learned representations on the following downstream tasks:

1. **Brain Age Prediction:** We used a combination of HCP-D and HCP-A. The age was z-score normalized.

2. **Sex classification:** We decided to use only HCP-A as the sex differences might be more obvious in the early developmental stage.
3. **Cognitive normal (CN) versus mild cognitive impairment (MCI) classification:** This was performed in the ADNI dataset. As 30% of the samples are MCI, we used a positive sample weighting in the cross-entropy loss.
4. **Prediction of Alzheimer’s disease assessment scale–cognitive subscale (ADAS-Cog):** The ADAS-Cog is a brief neuropsychological assessment to evaluate the severity of cognitive symptoms associated with dementia [40]. This task was performed in the ADNI dataset. The score was z-score normalized.

For all downstream tasks, we performed five-fold cross-validation. HCP-D and HCP-A [12] were divided into train/val/test splits with a 30/35/35 ratio. ADNI was divided using a 60/20/20 ratio, and the splits were stratified to maintain the same proportion of CN vs MCI subjects. For all downstream tasks, we used the pretrained model with D-RoPE and alternating masking, which achieved the best reconstruction performance. The hyperparameters used are detailed in Table 6 in the supplementary material.

We tested three approaches 1) fine-tuning the last encoder layer of the model without the decoder and adding a prediction head 2) linear probing on the frozen representations 3) an MLP head on top of the frozen representations.

All of these variants have as input the CLS token from the encoder (384-dimensional vector), while the input to the encoder are dMRI volumes cropped to a size of (192,224,56). As done previously [72], to show that our model can handle varying diffusion directions and b-values, we randomly sample 12 diffusion directions/b-value pairs per subject. This effectively corresponds to having a separate acquisition protocol per subject. A simpler comparison with just two different fixed protocols is included in Section 6.2 in the supplementary materials.

We compared our pretrained MAE with several baselines. We included a 3D ResNet18 (Which was used previously to regress age from DTI maps [25]) trained on the FA and MD maps in a fully supervised manner. The FA and MD maps were computed using DIPY [26] from the same 12 directions that were selected as input to our model. Additionally, we considered a standard ViT encoder (Adapted from [71]) with one of two inputs: the aforementioned FA and MD maps or the 12 randomly selected dMRI directions. For the latter, the different diffusion directions are considered as channels in the projection layer. The ViT included a 3D version of regular RoPE [60], which operates on the image space. The number of encoder layers (10) and embedding size (384) was chosen to match our pretrained MAE. Additionally, we included a comparison with typically used handcrafted features in dMRI [67]. We extracted the mean FA and MD values for different notable white matter tracts

(Tracts were obtained using TractSeg [65]) and concatenated the values into a 130-dimensional vector. Predictions were then obtained by either doing linear probing on top of this vector or adding an MLP head on top of it. Additionally, a multi-tissue multi-shell model [34] was used to estimate spherical harmonic (SH) coefficients of white matter up to the 4th order using MRtrix3 [62]. Rotation-invariant spherical harmonic (RISH) features were then computed by summing the squared magnitudes (power) of the SH coefficients within each harmonic order. We computed the mean value for each harmonic order along the Tractseg segmented white matter tracts and concatenated the values into a 165-dimensional vector. Predictions were then obtained by adding an MLP head on top.

5. Results

5.1. MAE Reconstruction

Exemplary reconstructions from the different masking strategies (Spatial, Diffusion, and Alternating) are shown in Figure 4 for one of the volumes with a b-value of 2000 s^2/mm . We compare the MAE with and without D-RoPE. Additional reconstructions for b-values of 1000 and 3000 s^2/mm are shown in the supplementary material (Figures 5 and 6). We observe a good reconstruction performance for most methods; however, the version without D-RoPE and diffusion space masking captures the structure well but not the general intensities. This is possibly due to the limited information it learns about the relationships between the different directions and b-values.

The quantitative evaluation of the different reconstructions is shown in terms of PSNR, SSIM, and FID in Table 1. The evaluation using the DTI metrics derived from the reconstructions is shown in Table 2. We observe that D-RoPE with diffusion masking and D-RoPE with alternating masking tend to provide better metrics overall (e.g. reductions of 37-40% in FID when comparing the later against MAE with Spatial masking). We also note that alternating masking helps the MAE without D-RoPE case to improve performance, possibly because it forces the model to combine more information about the spatial and diffusion spaces. Additionally, the low error in MD and relatively low error in FA suggest that the reconstructions are able to maintain the general microstructural features contained in the original dMRI volumes.

5.2. Downstream Evaluation

Table 3 summarizes the results for the age prediction and sex classification tasks. The methods are grouped into handcrafted (HC) features (FA and MD, and RISH by segmented tract), fully trained methods (3D ResNet and ViTs trained from scratch), and our partially trained methods (frozen features obtained from our pretrained encoder combined with

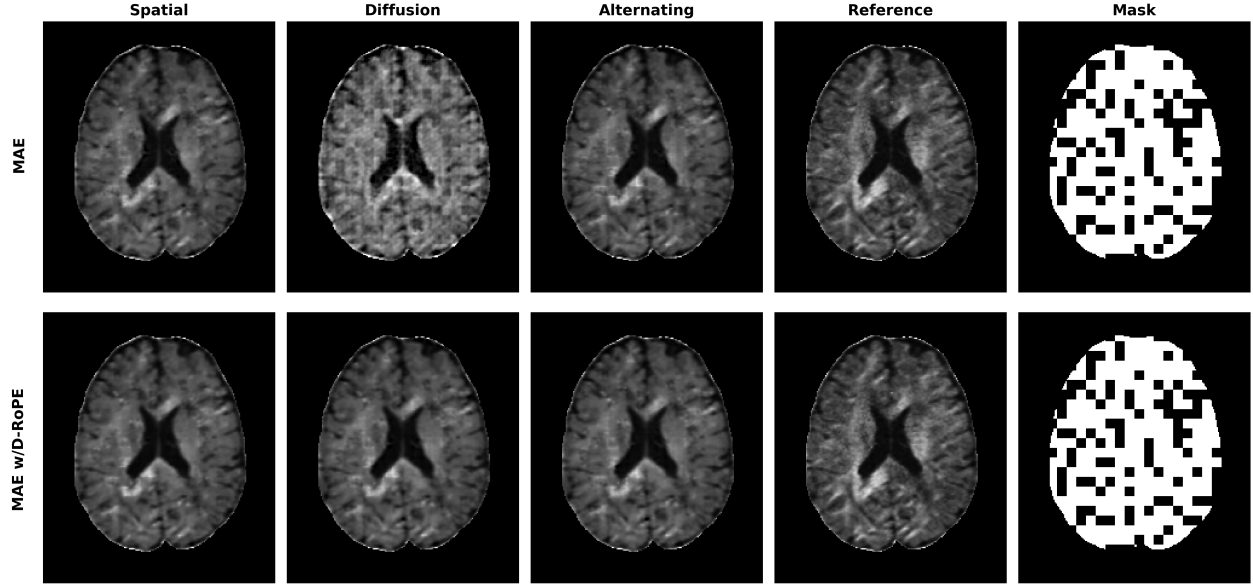


Figure 4. Qualitative evaluation of the reconstructions for a test subject for a b-value of $2000 \text{ s}^2/\text{mm}$. The reconstruction obtained with different masking strategies (Spatial, Diffusion, and Alternating) and with and without D-RoPE are shown in the three left-most columns. The reference ground-truth image and the spatial mask that was used are shown in the two right-most columns (white=masked). In general, a good reconstruction of the brain structures are observed. The MAE without D-RoPE case shows an erroneous contrast when only diffusion space masking is performed, possibly because of the lack of spatial information for a given diffusion direction.

either a linear layer or an MLP, and our pretrained encoder with an MLP head with only the last encoding layer finetuned). The mean and standard deviation across the five cross-validation folds are reported. We observe a good overall performance of most methods with a high correlation coefficient ($\rho > 0.89$), with the 3D ResNet and our finetuned pretrained model performing the best in terms of ρ and MSE. In the sex classification task, possibly due to the lower number of samples, a simple method such as using **frozen latents** from our pretrained encoder with a linear layer performs better than finetuning the model. Overall, our method provides the second-best performance, behind a fully supervised method, while we used frozen, not-fully-supervised features. The good performance of even just linear probing on our frozen latents suggests that the learned representations already encode some structural information that is useful for both of these tasks (Some evidence of this structure is shown in Figure 8 in the supplement).

Table 4 shows the results for the clinical tasks of cognitive score prediction and classification of cognitively normal (CN) versus mild cognitive impairment (MCI). For these tasks, the advantage of pretrained latent features becomes more apparent. Our frozen latents with an MLP head yield the best balanced accuracy (BA=64.7%) in the CN vs MCI classification and also the best overall performance in terms of correlation coefficient and MSE in the ADAS prediction task. In these two tasks, our frozen latents and the HC features outperform other methods that are more

data-intensive. The enhanced performance of our method with limited training data suggests that our pretrained encoder may have already internalized some microstructural patterns or features that are predictive of cognitive decline or disease transition. We also observe that our best performing method improves over ViT with standard RoPE applied on raw dMRI data. All metric comparisons between ViT with raw dMRI and our best performing method (except for MCI-AUROC and age-MSE) are statistically significant (paired t-test and Bonferroni correction).

6. Discussion and Conclusions

We have presented a novel transformer architecture designed to obtain general-purpose representations of diffusion MRI (dMRI). The model integrates mixed attention in both spatial and diffusion domains, along with a modified positional encoding specifically tailored to the geometry of the diffusion space (D-RoPE). Our findings indicate that the best performance is achieved by combining D-RoPE with alternating masking in the spatial and diffusion domains. As demonstrated in our reconstruction evaluation and ablation analysis, these components enable the model to encode relative information regarding different diffusion directions and b-values. Furthermore, the strong performance in FA and MD metrics derived from the reconstructed images suggests that sufficient microstructural information from the original images is encoded into the latent space, facilitating the re-

Table 1. Quantitative evaluation of the reconstructions obtained from different masking strategies (Spatial, Diffusion, and Alternating) and with and without D-RoPE in the test set of HCP-YA. The mean $\pm$ standard deviation is shown for Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). The value of the Fréchet Inception distance (FID) is shown in the last column. $\uparrow$: the higher, the better; $\downarrow$: the lower, the better. The best results per b-value are highlighted in bold.

	Method	Masking	PSNR ($\uparrow$)	SSIM ($\uparrow$)	FID ($\downarrow$)
b-value = 1000 s^2/mm	MAE	Spatial	14.54 $\pm$ 0.16	0.971 $\pm$ 0.003	13.37
	MAE	Diffusion	10.13 $\pm$ 0.26	0.875 $\pm$ 0.005	39.76
	MAE	Alternating	14.70 $\pm$ 0.17	0.971 $\pm$ 0.004	11.13
	MAE w/D-RoPE	Spatial	14.84 $\pm$ 0.17	0.973 $\pm$ 0.003	10.28
	MAE w/D-RoPE	Diffusion	14.86 $\pm$ 0.17	0.974 $\pm$ 0.003	10.73
	MAE w/D-RoPE	Alternating	14.94 $\pm$ 0.17	0.974 $\pm$ 0.003	8.36
b-value = 2000 s^2/mm	MAE	Spatial	14.91 $\pm$ 0.17	0.969 $\pm$ 0.003	13.49
	MAE	Diffusion	10.07 $\pm$ 0.28	0.844 $\pm$ 0.006	39.86
	MAE	Alternating	15.07 $\pm$ 0.18	0.968 $\pm$ 0.004	10.81
	MAE w/D-RoPE	Spatial	15.26 $\pm$ 0.17	0.972 $\pm$ 0.002	9.61
	MAE w/D-RoPE	Diffusion	15.31 $\pm$ 0.18	0.973 $\pm$ 0.002	9.99
	MAE w/D-RoPE	Alternating	15.37 $\pm$ 0.18	0.973 $\pm$ 0.002	8.05
b-value = 3000 s^2/mm	MAE	Spatial	15.13 $\pm$ 0.17	0.969 $\pm$ 0.003	14.97
	MAE	Diffusion	10.01 $\pm$ 0.30	0.819 $\pm$ 0.009	41.57
	MAE	Alternating	15.29 $\pm$ 0.18	0.968 $\pm$ 0.003	11.54
	MAE w/D-RoPE	Spatial	15.49 $\pm$ 0.17	0.971 $\pm$ 0.002	10.42
	MAE w/D-RoPE	Diffusion	15.54 $\pm$ 0.18	0.973 $\pm$ 0.002	10.63
	MAE w/D-RoPE	Alternating	15.61 $\pm$ 0.18	0.973 $\pm$ 0.002	8.95

Table 2. Evaluation of fractional anisotropy (FA) and mean diffusivity (MD) obtained from the reconstructions with and without D-RoPE in HCP-YA. For MD, the $\times 1000$ error is shown. The mean $\pm$ standard deviation of the absolute error is shown. $\downarrow$: the lower, the better. The best results are highlighted in bold.

Method	Masking	FA error ($\downarrow$)	MD error $\times 10^3$ ($\downarrow$)
MAE	Spatial	0.074 $\pm$ 0.013	0.098 $\pm$ 0.009
MAE	Diffusion	0.185 $\pm$ 0.012	0.233 $\pm$ 0.015
MAE	Alternating	0.080 $\pm$ 0.013	0.109 $\pm$ 0.008
MAE w/D-RoPE	Spatial	0.061 $\pm$ 0.007	0.104 $\pm$ 0.008
MAE w/D-RoPE	Diffusion	0.058 $\pm$ 0.007	0.099 $\pm$ 0.008
MAE w/D-RoPE	Alternating	0.055 $\pm$ 0.007	0.103 $\pm$ 0.008

construction of those features. We also note that, throughout our tests, random directions were employed, and varying numbers of directions were utilized between pre-trained and downstream tasks, underscoring the model’s ability to handle arbitrary diffusion acquisitions without the need to resample to a fixed acquisition scheme. However, a limitation of D-RoPE is that the distance we selected is not easily factorizable into keys and queries matrices, unlike standard RoPE, resulting in higher computational cost (For a sequence length l and head dimension h , the cost of D-RoPE

is $\mathcal{O}(lh^2)$ instead of $\mathcal{O}(lh)$ for regular RoPE).

The learned representations we obtained were rigorously evaluated across various downstream tasks using additional datasets. Both the pre-computed features with a lightweight architecture on top and a partially fine-tuned pre-trained model performed competitively or even better than the various baselines, particularly in tasks with limited training data. In particular, our method improves the performance over a standard ViT with regular RoPE with Raw dMRI as input. This suggests our joint spatial and diffusion modeling is needed in the case of varying diffusion protocols to more fully extract relevant microstructural information.

Several studies have demonstrated a statistical relationship between various dMRI microstructural metrics and age [19, 39, 41, 46]. Given this evidence, the strong performance of the different tested methods is not unexpected. The sex-classification task may be more challenging: dMRI microstructural sex differences have been reported during developmental ages [10, 42], but large-scale lifespan work [19] suggests that the nature and magnitude of these differences vary with age, which may complicate classification across different age groups.

Conversely, the downstream tasks we evaluated in the

Table 3. Evaluation of different methods (handcrafted features, 3D ResNet, ViT, and our partially supervised methods) on the age prediction and sex classification tasks in HCP-A and HCP-D. For age prediction, the correlation coefficient (ρ) and mean square error (MSE) are reported. For sex classification, accuracy (ACC) and area under the receiver operating curve (AUROC) are reported. $\uparrow$: the higher, the better, $\downarrow$: the lower, the better. The best result per metric is highlighted in bold, and the second best is underlined. Our proposed methods are highlighted in yellow. *HC Features means Hand-Crafted Features.

Method	Features	Prediction Head	Supervision	Age prediction		Sex classification	
				ρ ($\uparrow$)	MSE ($\downarrow$)	ACC ($\uparrow$)	AUROC ($\uparrow$)
HC Features	Tract FA, MD	Linear	Full	0.90 ± 0.01	0.20 ± 0.01	70.90 ± 2.60 %	0.77 ± 0.02
HC Features	Tract FA, MD	MLP	Full	0.90 ± 0.01	0.19 ± 0.02	56.90 ± 2.50 %	0.59 ± 0.03
HC Features	Tract based RISH	MLP	Full	0.61 ± 0.21	1.15 ± 1.24	61.70 ± 2.00 %	0.67 ± 0.03
3D ResNet	FA and MD maps	MLP	Full	0.96 ± 0.01	0.09 ± 0.02	76.00 ± 1.20 %	0.83 ± 0.04
ViT w/RoPE	FA and MD maps	MLP	Full	0.93 ± 0.01	<u>0.15 ± 0.02</u>	65.40 ± 3.70 %	0.72 ± 0.04
ViT w/RoPE	Raw dMRI	MLP	Full	0.89 ± 0.02	0.34 ± 0.11	65.10 ± 2.60 %	0.71 ± 0.02
Ours	Frozen latent	Linear	Head	0.89 ± 0.01	0.24 ± 0.02	<u>74.60 ± 2.90 %</u>	<u>0.82 ± 0.02</u>
Ours	Frozen latent	MLP	Head	0.90 ± 0.02	0.20 ± 0.03	59.30 ± 2.30 %	0.64 ± 0.02
Ours	Last layer finetuned	MLP	Partial	<u>0.94 ± 0.01</u>	<u>0.15 ± 0.03</u>	69.60 ± 3.10 %	0.77 ± 0.02

Table 4. Evaluation of different methods (handcrafted features, 3D ResNet, ViT, and our partially supervised methods) on the CN vs MCI classification and ADAS prediction tasks in ADNI. For CN vs MCI, balance accuracy (BA), area under the precision-recall curve (AUPRC), and area under the receiver operating curve (AUROC) are reported. For ADAS prediction, the correlation coefficient (ρ) and mean square error (MSE) are reported. $\downarrow$: the lower, the better. $\downarrow$: the lower, the better. The best result per metric is highlighted in bold, and the second best is underlined. Our proposed methods are highlighted in yellow. *HC Features means Hand-Crafted Features.

Method	Features	Prediction Head	Supervision	CN vs MCI classification			ADAS prediction	
				BA ($\uparrow$)	AUPRC ($\uparrow$)	AUROC ($\uparrow$)	ρ ($\uparrow$)	MSE ($\downarrow$)
HC Features	Tract FA, MD	Linear	Full	61.90 ± 7.30 %	0.53 ± 0.06	0.69 ± 0.06	0.06 ± 0.15	10.92 ± 4.08
HC Features	Tract FA, MD	MLP	Full	57.40 ± 5.40 %	0.44 ± 0.09	0.62 ± 0.04	<u>0.33 ± 0.09</u>	<u>0.81 ± 0.13</u>
HC Features	Tract based RISH	MLP	Full	53.90 ± 9.60 %	0.31 ± 0.07	0.52 ± 0.08	0.28 ± 0.14	0.93 ± 0.18
3D ResNet	FA and MD maps	MLP	Full	58.40 ± 8.30 %	<u>0.50 ± 0.10</u>	<u>0.68 ± 0.08</u>	<u>0.33 ± 0.08</u>	0.86 ± 0.14
ViT w/RoPE	FA and MD maps	MLP	Full	58.10 ± 6.60 %	0.47 ± 0.08	0.65 ± 0.06	0.25 ± 0.12	0.93 ± 0.15
ViT w/RoPE	Raw dMRI	MLP	Full	47.40 ± 5.30 %	0.29 ± 0.07	0.46 ± 0.13	-0.05 ± 0.18	1.08 ± 0.11
Ours	Frozen latent	Linear	Head	55.10 ± 3.50 %	0.44 ± 0.07	0.62 ± 0.11	0.18 ± 0.25	2.81 ± 1.12
Ours	Frozen latent	MLP	Head	64.70 ± 5.70 %	<u>0.50 ± 0.07</u>	0.67 ± 0.02	0.38 ± 0.14	0.77 ± 0.07
Ours	Last layer finetuned	MLP	Partial	60.90 ± 4.10 %	0.49 ± 0.09	0.66 ± 0.07	0.32 ± 0.19	0.88 ± 0.15

ADNI dataset (classification of MCI and prediction of the ADAS) are considerably more challenging. Implementing actual classification models to detect MCI in larger cohorts has proven to be more difficult [67] compared to classifying Alzheimer’s Disease (AD). Our pretrained latent representations demonstrate an improvement in balanced accuracy when compared to traditional handcrafted features (64.70% vs 61.90%) similar to those used in the literature [45, 67], indicating that the latent information likely encodes some microstructural patterns without the need to fit a biophysical model. Previous studies have shown correlations between dMRI metrics and various cognitive assessments [50, 70]. However, to the best of our knowledge, this is the first study to predict one of these cognitive assessments directly from dMRI volumes. The relatively high correlation coefficient we obtained ($\rho=0.38$) may indicate that the model can extract information from dMRI that is indicative of cognitive

trajectories in this particular cohort.

The representation learning framework for dMRI presented here is significant for several reasons. Our model is capable of harmonizing analyses across cohorts with varying acquisition protocols, serving as a stepping stone toward the development of a foundation model for dMRI, which can be achieved by increasing the diversity and size of pre-training data. Furthermore, the latent information or attention maps may reveal new microstructural biomarkers that link white matter integrity to cognitive domains or other diseases. Additionally, the approach presented here could potentially be adapted for other applications, such as super-resolution, by integrating additional components like latent diffusion models. Finally, the representations may also play a crucial role in a future multi-modal model that integrates different MRI modalities or other medical information.

Acknowledgments

This work was supported in part by National Institutes of Health grants AG089169, AG084471, DA057567 and AA021697, and the Stanford Institute for Human-Centered AI (HAI) Hoffman-Yee Award. A.C. receives research support from NIH grants R01 HL167974, R01HL169345, R01 AR077604, R01 EB002524, R01 AR079431, P41 EB 027060, P50 HD118632; Advanced Research Projects Agency for Health (ARPA-H) Biomedical Data Fabric (BDF) and Chatbot Accuracy and Reliability Evaluation (CARE) programs (contracts AY2AX000045 and 1AYSAX0000024-01); and the Medical Imaging and Data Resource Center (MIDRC), which is funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) under contract 75N92020C00021 and through ARPA-H.

References

- [1] MohammadHassan Abbasi, Gustavo Chau, and Ehsan Adeli. DTI Processing Pipeline: A reproducible framework for diffusion MRI preprocessing and quality control, 2025. Zenodo, doi: 10.5281/zenodo.17503171. [4](#)
- [2] Andrew L Alexander, Jung Hwan Lee, Mariana Lazar, and Aaron S Field. Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329, 2007. [1](#)
- [3] Jesper LR Andersson and Stamatios N Sotiropoulos. An integrated approach to correction for off-resonance effects and subject movement in diffusion mr imaging. *NeuroImage*, 125:1063–1078, 2016. [4](#)
- [4] Jesper LR Andersson, Stefan Skare, and John Ashburner. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage*, 20(2):870–888, 2003. [4](#)
- [5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. [2](#)
- [6] Yaniv Assaf and Ofer Pasternak. Diffusion mr imaging. *NeuroImage*, 45:S27–S38, 2008. [1](#)
- [7] Brian B Avants, Charles L Epstein, Murray Grossman, and James C Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008. [4](#)
- [8] Carlo Alberto Barbano, Matteo Brunello, Benoit Dufumier, and Marco Grangetto. Anatomical foundation models for brain mris. *arXiv preprint arXiv:2408.07079*, 2024. [2](#)
- [9] Peter J Basser, James Mattiello, and Denis LeBihan. Mr diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267, 1994. [1](#)
- [10] Sebastian M Benavidez, Zvart Abaryan, Gaon S Kim, Emily Laltoo, James T McCracken, Paul M Thompson, and Katherine E Lawrence. Sex differences in the brain’s white matter microstructure during development assessed using advanced diffusion mri models. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–7. IEEE, 2024. [7](#)
- [11] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. [2](#), [3](#)
- [12] Susan Y Bookheimer, David H Salat, Melissa Terpstra, Beau M Ances, Deanna M Barch, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Diaz-Santos, Jennifer Stine Elam, et al. The lifespan human connectome project in aging: an overview. *Neuroimage*, 185:335–348, 2019. [4](#), [5](#)
- [13] Marco Bozzali, Andrea Falini, Massimo Franceschi, Mara Cercignani, Maddalena Zuffi, Giuseppe Scotti, Giancarlo Comi, and Massimo Filippi. White matter damage in alzheimer’s disease assessed in vivo using diffusion tensor magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 72(6):742–746, 2002. [1](#)
- [14] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *Advances in neural information processing systems*, 34:19594–19607, 2021. [2](#)
- [15] Josue Ortega Caro, Antonio H de O Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. Brainlm: A foundation model for brain activity recordings. *bioRxiv*, pages 2023–09, 2023. [2](#)
- [16] Geng Chen, Yoonmi Hong, Yongqin Zhang, Jaeil Kim, Khoi Minh Huynh, Jiquan Ma, Weili Lin, Dinggang Shen, Pew-Thian Yap, and UNC/UMN Baby Connectome Project Consortium. Estimating tissue microstructure with under-sampled diffusion data via graph convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 280–290. Springer, 2020. [2](#)
- [17] Geng Chen, Haotian Jiang, Jiannan Liu, Jiquan Ma, Hui Cui, Yong Xia, and Pew-Thian Yap. Hybrid graph transformer for tissue microstructure estimation with undersampled diffusion mri data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 113–122. Springer, 2022. [2](#)
- [18] Junbo Chen, Vara Lakshmi Bayanagari, Sohae Chung, Yao Wang, and Yvonne W Lui. Deep learning with diffusion mri as in vivo microscope reveals sex-related differences in human white matter microstructure. *Scientific reports*, 14(1): 9835, 2024. [2](#)
- [19] Stefania Conte, Dabriel Zimmerman, and John E Richards. White matter trajectories over the lifespan. *Plos one*, 19(5): e0301520, 2024. [7](#)
- [20] Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Chong, Fang Ji, Nathanael Tong, Christopher Chen, and Juan Helen Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073, 2024. [2](#)
- [21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

- Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3
- [22] Marta Drake-Pérez, Jose Boto, Aikaterini Fitsiori, Karl Lovblad, and Maria Isabel Vargas. Clinical applications of diffusion weighted imaging in neuroradiology. *Insights into imaging*, 9(4):535–547, 2018. 1
- [23] Jennifer Stine Elam, Matthew F Glasser, Michael P Harms, Stamatios N Sotiropoulos, Jesper LR Andersson, Gregory C Burgess, Sandra W Curtiss, Robert Oostenveld, Linda J Larson-Prior, Jan-Mathijs Schoffelen, et al. The human connectome project: a retrospective. *NeuroImage*, 244:118543, 2021. 4
- [24] Massimo Filippi, Mara Cercignani, Matilde Inglese, Mark A Horsfield, and Giancarlo Comi. Diffusion tensor magnetic resonance imaging in multiple sclerosis. *Neurology*, 56(3):304–311, 2001. 1
- [25] Chenyu Gao, Michael E Kim, Ho Hin Lee, Qi Yang, Nazirah Mohd Khairi, Praitayini Kanakaraj, Nancy R Newlin, Derek B Archer, Angela L Jefferson, Warren D Taylor, et al. Predicting age from white matter diffusivity with residual learning. *ArXiv*, pages arXiv–2311, 2024. 5
- [26] Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, Ian Nimmo-Smith, and Dipy Contributors. Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics*, 8:8, 2014. 4, 5
- [27] Eric K Gibbons, Kyler K Hodgson, Akshay S Chaudhari, Lorie G Richards, Jennifer J Majersik, Ganesh Adluru, and Edward VR DiBella. Simultaneous noddif and gfa parameter map generation from subsampled q-space imaging using deep learning. *Magnetic resonance in medicine*, 81(4):2399–2411, 2019. 2
- [28] Matthew F Glasser, Stamatios N Sotiropoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013. 4
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *corr abs/1512.03385 (2015)*, 2015. 4
- [30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. “masked autoencoders are scalable vision learners”. *arXiv preprint arXiv:2111.06377*, 2021. 2
- [31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4
- [32] Shuo Huang, Lujia Zhong, and Yonggang Shi. Autoencoder for 4-dimensional fiber orientation distributions from diffusion mri. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2025. 2
- [33] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012. 4
- [34] Ben Jeurissen, Jacques-Donald Tournier, Thijs Dhollander, Alan Connelly, and Jan Sijbers. Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion mri data. *NeuroImage*, 103:411–426, 2014. 5
- [35] Heidi Johansen-Berg and Timothy EJ Behrens. Diffusion mri in psychiatry. *Neuroimaging in psychiatry*, pages 51–68, 2009. 1
- [36] Derek K Jones, Thomas R Knösche, and Robert Turner. White matter integrity, fiber count, and other fallacies: the do’s and don’ts of diffusion mri. *Neuroimage*, 73:239–254, 2013. 1
- [37] Davood Karimi. An approach to building foundation models for brain image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 421–431. Springer, 2024. 2
- [38] Davood Karimi and Ali Gholipour. Diffusion tensor estimation with transformer neural networks. *Artificial intelligence in medicine*, 130:102330, 2022. 2
- [39] P Kochunov, DE Williamson, J Lancaster, P Fox, J Cornell, J Blangero, and DC Glahn. Fractional anisotropy of water diffusion in cerebral white matter across the lifespan. *Neurobiology of aging*, 33(1):9–20, 2012. 7
- [40] Jacqueline K Kueper, Mark Speechley, and Manuel Montero-Odasso. The alzheimer’s disease assessment scale–cognitive subscale (adas-cog): modifications and responsiveness in pre-dementia populations. a narrative review. *Journal of Alzheimer’s Disease*, 63(2):423–444, 2018. 5
- [41] Rajesh Kumar, Alexa S Chavez, Paul M Macey, Mary A Woo, and Ronald M Harper. Brain axial and radial diffusivity changes with age and gender in healthy adults. *Brain research*, 1512:22–36, 2013. 7
- [42] Katherine E Lawrence, Zvart Abaryan, Emily Laltoo, Leanna M Hernandez, Michael J Gandal, James T McCracken, and Paul M Thompson. White matter microstructure shows sex differences in late childhood: Evidence from 6797 children. *Human Brain Mapping*, 44(2):535–548, 2023. 7
- [43] Zhiyuan Liu, Zekun Jiang, Li Meng, Jun Yang, Ying Liu, Yingying Zhang, Haiqin Peng, Jiahui Li, Gang Xiao, Zijian Zhang, et al. Handcrafted and deep learning-based radiomic models can distinguish gbm from brain metastasis. *Journal of Oncology*, 2021(1):5518717, 2021. 2
- [44] Matthew Lyon, Paul Armitage, and Mauricio A Álvarez. Spatio-angular convolutions for super-resolution in diffusion mri. *Advances in Neural Information Processing Systems*, 36:12457–12475, 2023. 2
- [45] Tommaso Maggipinto, Roberto Bellotti, Nicola Amoroso, Domenico Diacono, Giacinto Donvito, Eufemia Lella, Alfonso Monaco, Marzia Antonella Scelsi, Sabina Tangaro, Alzheimer’s Disease Neuroimaging Initiative, et al. Dti measurements for alzheimer’s classification. *Physics in Medicine & Biology*, 62(6):2361, 2017. 8
- [46] Stephanie Matijevic and Lee Ryan. Tract specificity of age effects on diffusion tensor imaging measures of white matter health. *Frontiers in Aging Neuroscience*, 13:628865, 2021. 7

- [47] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 4
- [48] Susumu Mori, Brian J Crain, Venkateswaran P Chacko, and Peter CM Van Zijl. Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. *Annals of neurology*, 45(2):265–269, 1999. 1
- [49] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer’s disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005. 4
- [50] Talia M Nir, Neda Jahanshad, Julio E Villalon-Reina, Arthur W Toga, Clifford R Jack, Michael W Weiner, Paul M Thompson, Alzheimer’s Disease Neuroimaging Initiative (ADNI, et al. Effectiveness of regional dti measures in distinguishing alzheimer’s disease, mci, and normal aging. *NeuroImage: clinical*, 3:180–195, 2013. 8
- [51] Dmitry S Novikov, Valerij G Kiselev, and Sune N Jespersen. Quantifying brain microstructure with diffusion mri: Theory and parameter estimation. *NMR in Biomedicine*, 32(4): e3998, 2019. 1
- [52] Sophie Ostmeier, Brian Axelrod, Michael E Moseley, Akshay Chaudhari, and Curtis Langlotz. Liere: Generalizing rotary position encodings. *arXiv preprint arXiv:2406.10322*, 2(4), 2024. 2
- [53] Carlo Pierpaoli and et al. Diffusion tensor imaging in aging and dementia. *NMR in Biomedicine*, 15(7-8):456–467, 2001. 1
- [54] Ning Qiang, Qinglin Dong, Fangfei Ge, Hongtao Liang, Bao Ge, Shu Zhang, Yifei Sun, Jie Gao, and Tianming Liu. Deep variational autoencoder for mapping functional brain networks. *IEEE Transactions on Cognitive and Developmental Systems*, 13(4):841–852, 2020. 2
- [55] Sumeet Shinde, Shweta Prasad, Yash Saboo, Rishabh Kaushick, Jitender Saini, Pramod Kumar Pal, and Madhura Ingahlalikar. Predictive markers for parkinson’s disease using deep neural nets on neuromelanin sensitive mri. *NeuroImage: Clinical*, 22:101748, 2019. 2
- [56] Vishwa Mohan Singh, Alberto Gaston Villagran Asiares, Luisa Sophie Schuhmacher, Kate Rendall, Simon Weißbrod, David Rügamer, and Inga Körte. An interpretable representation learning approach for diffusion tensor imaging. *arXiv preprint arXiv:2505.19110*, 2025. 2
- [57] Stephen M Smith, Mark Jenkinson, Mark W Woolrich, Christian F Beckmann, Timothy EJ Behrens, Heidi Johansen-Berg, Peter R Bannister, Marilena De Luca, Ivana Drobniak, David E Flitney, et al. Advances in functional and structural mr image analysis and implementation as fsl. *Neuroimage*, 23:S208–S219, 2004. 4
- [58] Leah H Somerville, Susan Y Bookheimer, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Dapretto, Jennifer Stine Elam, Michael S Gaffrey, Michael P Harms, Cynthia Hodge, et al. The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage*, 183: 456–468, 2018. 4
- [59] Edward O. Stejskal and John E. Tanner. Spin diffusion measurements: Spin echoes in the presence of a time-dependent field gradient. *The Journal of Chemical Physics*, 42(1):288–292, 1965. 2
- [60] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 2, 3, 5
- [61] Divyanshu Tak, Biniam A Garomsa, Tafadzwa L Chaunzwa, Anna Zapaishchykova, Juan Carlos Climent Pardo, Zezhong Ye, John Zielke, Yashwanth Ravipati, Sri Vajapeyam, Maryam Mahootiha, et al. A foundation model for generalized brain mri analysis. *medRxiv*, 2024. 2
- [62] J-Donald Tournier, Robert Smith, David Raffelt, Rami Tabbara, Thijs Dhollander, Maximilian Pietsch, Daan Christiaens, Ben Jeurissen, Chun-Hung Yeh, and Alan Connelly. Mrtrix3: A fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage*, 202:116137, 2019. 5
- [63] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2
- [64] Shuangxing Wang, Lihui Wang, Ying Cao, Zeyu Deng, Chen Ye, Rongpin Wang, Yuemin Zhu, and Hongjiang Wei. Self-supervised arbitrary-scale super-angular resolution diffusion mri reconstruction. *Medical Physics*, 52(5):2976–2998, 2025. 2
- [65] Jakob Wasserthal, Peter Neher, and Klaus H Maier-Hein. Tractseg-fast and accurate white matter tract segmentation. *Neuroimage*, 183:239–253, 2018. 4, 5
- [66] Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? *arXiv preprint arXiv:2502.05173*, 2025. 2
- [67] Junhao Wen, Jorge Samper-González, Simona Bottani, Alexandre Routier, Ninon Burgos, Thomas Jacquemont, Sabrina Fontanella, Stanley Durrleman, Stephane Epelbaum, Anne Bertrand, et al. Reproducible evaluation of diffusion mri features for automatic classification of patients with alzheimer’s disease. *Neuroinformatics*, 19(1):57–78, 2021. 5, 8
- [68] Taohui Xiao, Jian Cheng, Wenxin Fan, Enqing Dong, and Shanshan Wang. Samrobnoddi: q-space sampling-augmented continuous representation learning for robust and generalized noddi. *Physics in Medicine and Biology*, 2024. 2
- [69] Junqing Yang, Haotian Jiang, Tewodros Tassew, Peng Sun, Jiquan Ma, Yong Xia, Pew-Thian Yap, and Geng Chen. Towards accurate microstructure estimation via 3d hybrid graph transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 25–34. Springer, 2023. 2
- [70] Artemis Zavaliangos-Petropulu, Talia M Nir, Sophia I Thomopoulos, Robert I Reid, Matt A Bernstein, Bret Borowski,

Clifford R Jack Jr, Michael W Weiner, Neda Jahanshad, and Paul M Thompson. Diffusion mri indices and their relation to cognitive impairment in brain aging: the updated multi-protocol approach in adni3. *Frontiers in Neuroinformatics*, 13:2, 2019. [8](#)

- [71] Tianshu Zheng, Guohui Yan, Haotian Li, Weihao Zheng, Wen Shi, Yi Zhang, Chuyang Ye, and Dan Wu. A microstructure estimation transformer inspired by sparse representation for diffusion mri. *Medical Image Analysis*, 86: 102788, 2023. [2](#), [5](#)
- [72] Fangrong Zong, Zaimin Zhu, Jiayi Zhang, Xiaofeng Deng, Zhuangzhuang Li, Chuyang Ye, and Yong Liu. Attention-based q-space deep learning generalized for accelerated diffusion magnetic resonance imaging. *IEEE Journal of Biomedical and Health Informatics*, 2024. [2](#), [5](#)