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

Supplementary Material

6.1. Training and fine-tuning parameters

The parameters used for the pretraining stage of the MAE are detailed in Table 5, while those used for finetuning the last layer and the head of the MAE for the different downstream tasks are shown in Table 6.

Table 5. Parameters using for the pre-training stage of the MAE.

Hyperparameter	Value
Number of transformer encoding Layers	10
Number of transformer decoding Layers	3
Number of convolutional decoding Layers	3
Latent dimension	384
Number of heads	3
Convolutional kernel size	3
optimizer	AdamW
optimizer momentum	β_1=0.9, β_2=0.999
learning rate schedule	cosine
warmup epochs	40
start learning rate	5.00E-05
final learning rate	1.00E-06
weight decay schedule	cosine
initial weight decay	0.04
final weight decay	0.4
batch size	4
epochs	300

Table 6. Parameters using for the partial finetuning of the MAE in different downstream tasks.

	Task			
hyperparameter	Age prediction	Sex Classification	MCI classification	ADAS prediction
optimizer	AdamW	AdamW	AdamW	AdamW
optimizer momentum	$\beta_1 = 0.9$, $\beta_2 = 0.999$	$\beta_1 = 0.9$, $\beta_2 = 0.999$	$\beta_1 = 0.9$, $\beta_2 = 0.999$	$\beta_1 = 0.9$, $\beta_2 = 0.999$
learning rate schedule	Constant	cosine	Constant	cosine
start learning rate	5.00E-05	1.00E-05	5.00E-05	5.00E-05
final learning rate	5.00E-05	1.00E-06	5.00E-05	1.00E-06
weight decay schedule	Constant	Constant	Constant	cosine
initial weight decay	0.01	0.01	0.05	0.05
final weight decay	0.01	0.01	0.05	0.01
batch size	6	6	6	6
epochs	50	50	50	20
LoRA used?	No	No	Yes	Yes
rank (LoRA)	-	-	12	12
alpha (LoRA)	-	-	12	12

6.2. Additional evaluations

Exemplary reconstructions from the different masking strategies (Spatial, Diffusion and Alternating) are shown in Figures 5 and 6 for volumes with a b-values of 1000 s^2/mm and 3000 s^2/mm, respectively. At 3000 s^2/mm, it is clearly observed how the version without D-RoPE and diffusion space masking captures well the structure but not the general intensity.

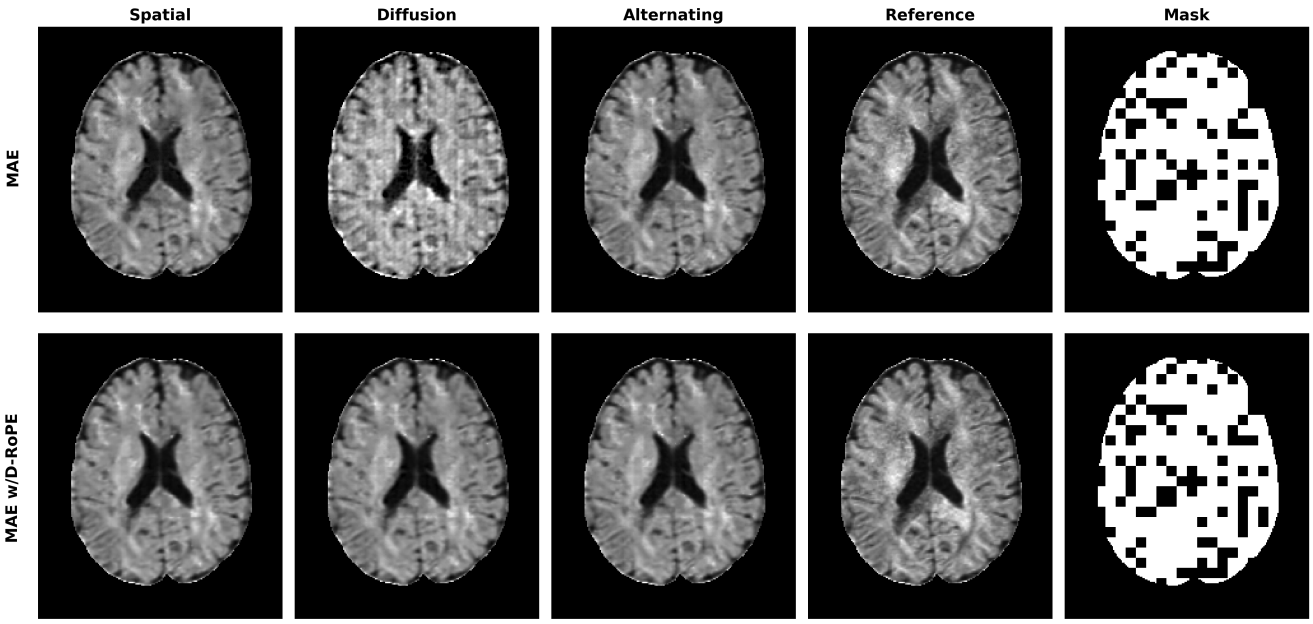

Figure 5. Qualitative evaluation of the reconstructions for a test subject for a b-value of 1000 s^2/mm. The reconstruction obtained from different masking strategies (Spatial, Diffusion and Alternating) and with and without D-RoPE are shown in the three left-most columns. The reference ground-truth image and the spatial mask that was used are shown in the two right-most columns.

Figure 7 shows additional ablations with different patch sizes (Top) and at different stages of pretraining (Bottom). We evaluated patch sizes of (16,16,4), (12,14,4) and (8,8,4) and checkpoints of the model at 50, 100, 150, 200, 250 and 300 epochs in the pretraining stage. All other parameters are kept as in the baseline model (Table 5). The evaluation was obtained with linear probing on top of the frozen latents as described in Section 4.3 and the plotted points correspond to the mean across the five-fold cross-validation setup. We observe a clear performance improvement in all downstream tasks as we utilize smaller patch sizes (e.g an improvement of approximately 5% in sex classification going from the largest patch size to the smallest patch size). The results for different pretraining epochs vary more in their optimal points from one downstream task to another, with at least improvement in all metrics observed when going from 50 to 100 epochs.

To further show the effectiveness of our method in handling different protocols, we designed an experiment where two different protocols of 12 random b-values and b-vectors are used, and latent representations are obtained with our pretrained model. A linear classifier was trained on Protocol 1 and then tested on either Protocol 1 or Protocol 2. The results are shown in Table 7. We observe minimal variation between the results obtained from the representations of both protocols, suggesting that the latent representations are consistent across protocols and that the model is able to handle arbitrary protocols.

Table 7. Comparison of linear probing on top of latent representations trained on protocol 1 and tested on this and a second different protocol. Minimal changes in performance are observed.

		Age prediction		Sex classification	
Trained on	Tested on	ρ (↑)	MSE (↓)	ACC (↑)	AUROC (↑)
Protocol 1	Protocol 1	0.895 ± 0.010	0.22 ± 0.02	75.0 ± 1.2 %	0.827 ± 0.015
Protocol 1	Protocol 2	0.897 ± 0.007	0.21 ± 0.02	75.4 ± 2.0 %	0.828 ± 0.017

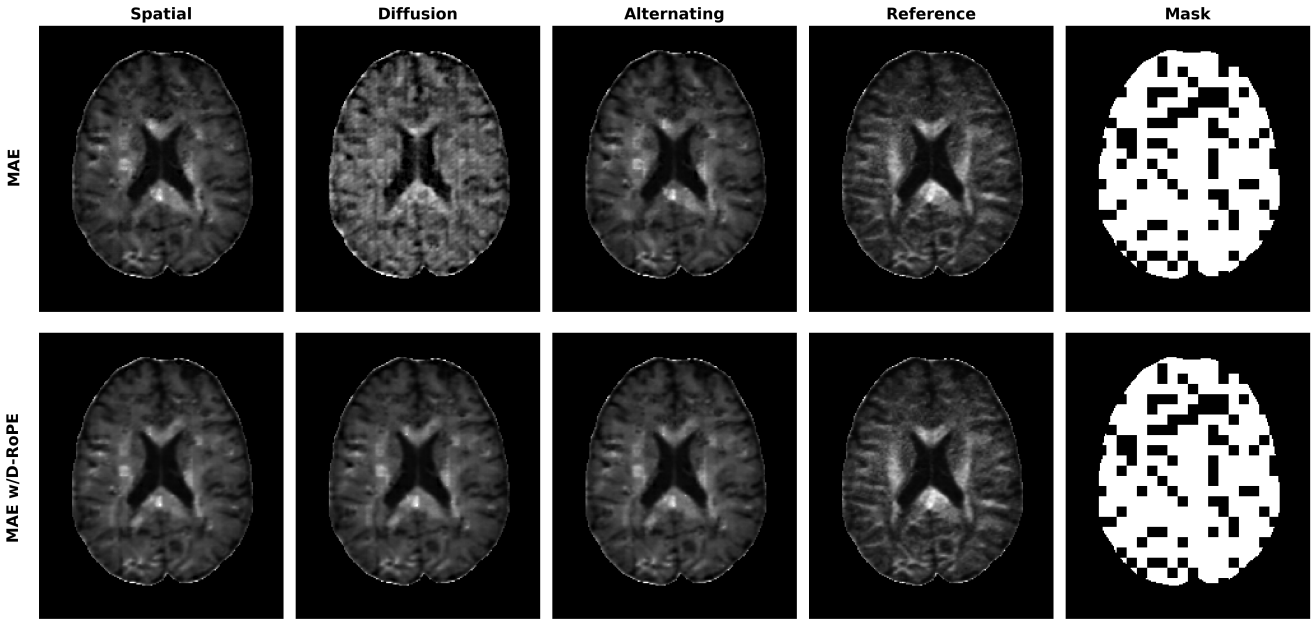

Figure 6. Qualitative evaluation of the reconstructions for a test subject for a b-value of 3000 s^2/mm. The reconstruction obtained from different masking strategies (Spatial, Diffusion and Alternating) and with and without D-RoPE are shown in the three left-most columns. The reference ground-truth image and the spatial mask that was used are shown in the two right-most columns.

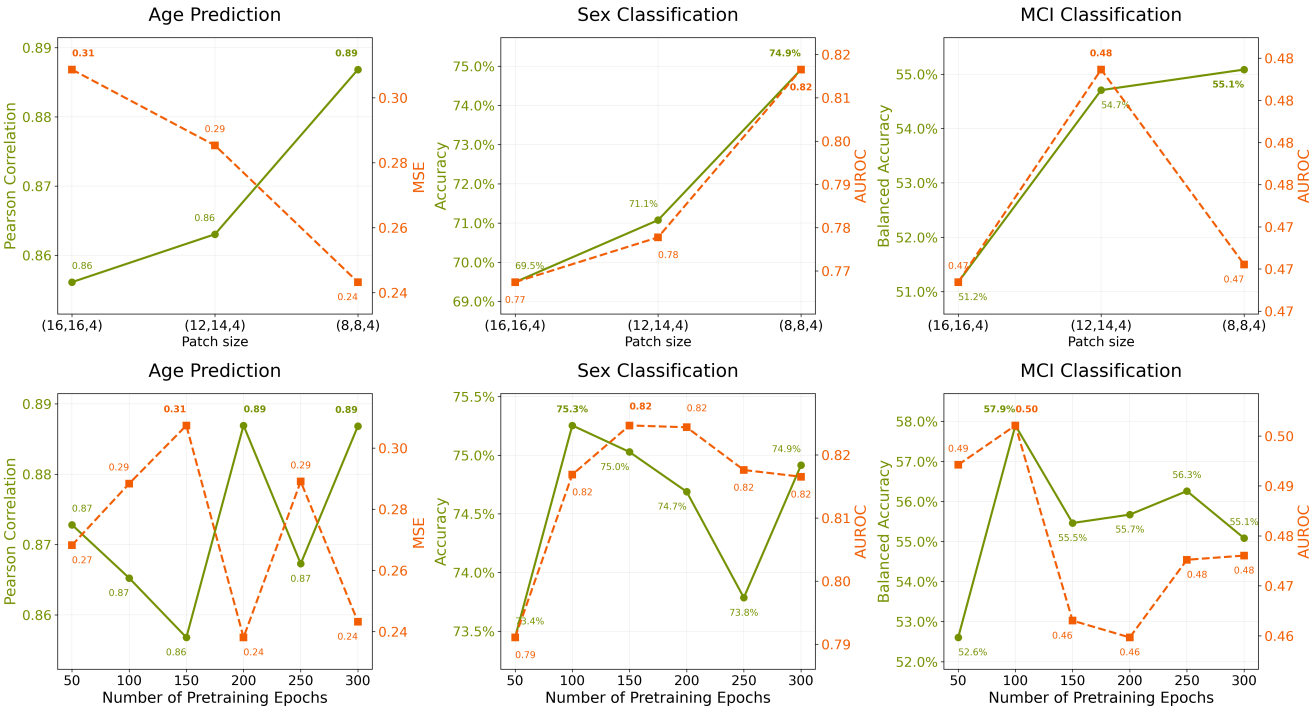

Figure 7. Linear probing performance of the frozen latents with different patch sizes (Top) and at different stages of pretraining (Bottom). A clear performance improvement in all downstream tasks is observed as one goes to smaller patch sizes. The optimal number of pretraining epochs seems to vary depending on the downstream task.

6.3. Interpretability

To explore the representations obtained with the pretrained model, Figure 8 shows UMAP [47] projections for the HCP-A samples colored by sex on the left and combined HCP-A and HCP-D colored by age on the right. We observe certain qualitative structure of the latent space even without any finetuning. Female subjects appear to be concentrated towards the lower part of the manifold, and there is a gradient from younger to older subjects going upwards in the manifold.

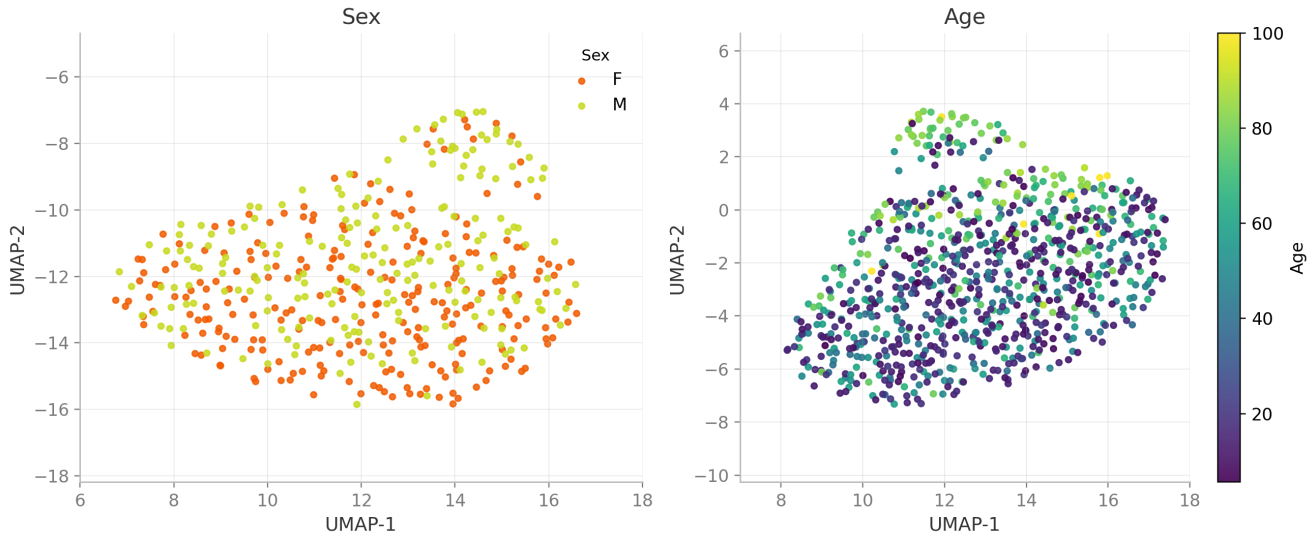

Figure 8. Two dimensional UMAP projections for the HCP-A samples colored by sex on the left and combined HCP-A and HCP-D samples colore by age on the right. We observed a structure of the latent space in which female subjects appear to be concentrated towards the lower part of the manifold, and in which there is a gradient from younger to older subjects going upwards in the manifold.