

# Morphology-preserving Autoregressive 3D Generative Modelling of the Brain

Petru-Daniel Tudosiu<sup>@\*1[0000-0001-6435-5079]</sup>, Walter Hugo Lopez Pinaya<sup>1[0000-0002-2795-9209]</sup>, Mark S. Graham<sup>1[0000-0002-4170-1095]</sup>, Pedro Borges<sup>1[0000-0001-5357-1673]</sup>, Virginia Fernandez<sup>1[0000-0001-5984-197X]</sup>, Dai Yang<sup>2</sup>, Jeremy Appleyard<sup>2</sup>, Guido Novati<sup>#3</sup>, Disha Mehra<sup>2</sup>, Mike Vella<sup>#4</sup>, Parashkev Nachev<sup>5[0000-0002-2718-4423]</sup>, Sebastien Ourselin<sup>1[0000-0002-5694-5340]</sup>, and Jorge Cardoso<sup>1[0000-0003-1284-2558]</sup>

<sup>1</sup> Department of Biomedical Engineering, School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK

<sup>2</sup> NVIDIA

<sup>3</sup> DeepMind

<sup>4</sup> Oxford Nanopore Technologies, Gosling Building, Oxford Science Park, Edmund Halley Rd, Littlemore, Oxford OX4 4DQ, UK

<sup>5</sup> Queen Square Institute of Neurology, University College London, London, UK

<sup>@</sup> Corresponding Author

<sup>#</sup> Work done while at NVIDIA

**Abstract.** Human anatomy, morphology, and associated diseases can be studied using medical imaging data. However, access to medical imaging data is restricted by governance and privacy concerns, data ownership, and the cost of acquisition, thus limiting our ability to understand the human body. A possible solution to this issue is the creation of a model able to learn and then generate synthetic images of the human body conditioned on specific characteristics of relevance (e.g., age, sex, and disease status). Deep generative models, in the form of neural networks, have been recently used to create synthetic 2D images of natural scenes. Still, the ability to produce high-resolution 3D volumetric imaging data with correct anatomical morphology has been hampered by data scarcity and algorithmic and computational limitations. This work proposes a generative model that can be scaled to produce anatomically correct, high-resolution, and realistic images of the human brain, with the necessary quality to allow further downstream analyses. The ability to generate a potentially unlimited amount of data not only enables large-scale studies of human anatomy and pathology without jeopardizing patient privacy, but also significantly advances research in the field of anomaly detection, modality synthesis, learning under limited data, and fair and ethical AI. Code and trained models are available at: <https://github.com/AmigoLab/SynthAnatomy>.

**Keywords:** Transformers · VQ-VAE · Generative Modelling · Neuroimaging · Neuromorphology.

## 1 Introduction

Current advances in the application of deep learning (DL) in medical imaging were driven by substantial initiatives and challenges such as UK Biobank (UKB)[1], Alzheimer’s Disease Neuroimaging Initiative (ADNI)[2], and the Medical Segmentation Decathlon[3]. However, these are relatively small compared to computer vision datasets. Owing to the lack of access to sufficient data due to privacy concerns, medical imaging data is not fully leveraging DL’s full potential and this hinders its translation from research to the clinical environment. State-of-the-art (SOTA) algorithms rely on a handful of highly curated datasets which could lead to biases due to imbalanced demographics or acquisition parameters, that may negatively affect their performance for certain populations. A solution to this problem could come from the generative modelling of the underlying available data to balance the prevalence of confounding variables in the training dataset.

While semi-supervised 3D generative modelling of the brain has been steadily explored and improved [4, 5, 6], progress in unsupervised generative modelling has been more limited. Generative Adversarial Network (GAN) based approaches, which suffer from memory constraints and stability issues, have mostly been trained on low-resolution 3D images [7, 8, 9], having only recently been able to synthesise full resolution images via learning partial sub-volumes [10]. Whereas previous methods quantify sample diversity using classic metrics such as Multi-Scale Structural Similarity Index (MS-SSIM) [11], distribution alignment via Fréchet Inception Distance (FID) [12] and Maximum Mean Discrepancy (MMD) [13], none have quantified if the generated data preserves the morphological characteristics of the data – crucial if we are to use such methods.

Recently, autoregressive models have achieved SOTA results synthesising high resolution natural images [14, 15, 16]. This was accomplished by employing a compression model, namely a Vector Quantised-Variational Autoencoder (VQ-VAE) [17, 14], to project the images into a discrete latent representation where the images’ likelihood becomes tractable. An attention-based Transformer network [18, 19] is then used to model the product of conditional distributions by maximising the expected log-likelihood of the training data.

Following [20] as part of the Synthetic Data Desiderata, a good synthetic dataset should share many if not all statistical properties of the real dataset. One such property, if not the most important, of synthetic structural medical images is their morphological correctness. Covariates of interest such as demographic and pathological ones determine the phenotype of each subject which in turn contributes to the population-level morphological statistics. Without it, any development done on the synthetic data as part of the Train on Synthetic, Test on Real [21] paradigm could suffer from higher domain distribution shifts slowing down the development. Furthermore, without morphological assessment, any hypothesis tested on the synthetic data would be rendered highly uncertain.

In this study, we scale and optimise VQ-VAE and Transformer models for high-resolution volumetric data, aiming to learn the data distribution of both radiologically healthy and pathological brains. A thorough morphological evalu-

ation is employed by using Voxel-Based Morphometry (VBM) [22] and volumetric analysis using Geodesic Information Flows (GIF) [23], demonstrating that synthetic data generated by the proposed model preserves the morphological characteristics and phenotype of the data.

## 2 Background

Our model is based on the two-stage architecture introduced by [17, 14] and extended by [15], where a VQ-VAE model is used to project a high-resolution image into a compressed latent representation and a transformer is trained to maximize the likelihood of the flattened representations.

### 2.1 VQ-VAE

The VQ-VAE [17, 14] is comprised of an encoder  $E$  that projects the input image  $\mathbf{x} \in \mathbb{R}^{H \times W \times D}$  to a latent representation space  $\hat{\mathbf{z}} \in \mathbb{R}^{h \times w \times d \times n_z}$  where  $n_z$  is the latent embedding vector’s dimensionality. Afterwards, an element-wise quantization is done for each spatial code  $\hat{\mathbf{z}}_{ijk} \in \mathbb{R}^{n_z}$  onto its nearest vector  $e_k \in \mathbb{R}^{n_z}, k \in 1, \dots, K$  from a codebook, where  $K$  denotes the vocabulary size of the codebook, obtaining  $\hat{\mathbf{z}}_q$ . The codebook’s elements are learned in an online manner, together with the other model’s parameters. Based on the quantized latent space, a decoder  $G$  tries to reconstruct the observations  $\hat{\mathbf{x}} \in \mathbb{R}^{H \times W \times D}$ . By replacing each of the codebook elements vector  $\hat{\mathbf{z}}_q \in \mathbb{R}^{h \times w \times d \times n_z}$  with their associated index  $k$ , the latent discrete representation is obtained.

### 2.2 Transformer

Transformers models and their associated self-attention mechanisms can capture the interactions between inputs regardless of their relative positioning. Due to this, the attention mechanism scales quadratically with the size of the input sequence. Since the VQ-VAE’s latent discrete representation when applied to volumetric medical data is 3D and thus large in scale, standard transformers do not scale to the necessary sequence length. Recently, multiple advances have made Transformers more efficient [24]; models such as the Performer, with its FAVOR+ linear scaling attention approximation [19] offers a good compromise between accurately modelling long-sequences while preserving a reasonable computational complexity [24]. Thus the Performer is used to model the latent sequences; by minimizing the conditional distribution of codebook indices  $p(s_i) = p(s_i | s_{<i})$  on the flattened 1D sequences of the 3D latent discrete representations, the data log-likelihood is maximized in an autoregressive fashion.

## 3 Methods

### 3.1 Descriptive Quantization for Transformer Usage

To create a Transformer-based generative model of the brain, the image volume needs to be transformed into a 1D sequence of tokens. To achieve this, a VQ-VAE

model that reduces the overall spatial size by a factor of 4096, allowing an input image of size  $X$  to be represented by a sequence of 1400 tokens. This 1400-long token sequence is learnt in an online fashion together with the VQ-VAE model by using the Exponential Moving Average (EMA) algorithm [17, 14] as per Eq. 1.

$$\mathcal{L}_{VQ-VAE}(\mathbf{x}, G(\hat{\mathbf{z}}_{\mathbf{q}})) = \mathcal{L}_{Rec} + \mathcal{L}_{Adv} + \|sg[E(\mathbf{x})] - \hat{\mathbf{z}}_{\mathbf{q}}\|_2^2 + \beta \|sg[\hat{\mathbf{z}}_{\mathbf{q}}] - E(\mathbf{x})\|_2^2 \quad (1)$$

$$N_i^{(t)} := N_i^{(t-1)} * \gamma + n_i^{(t)}(1 - \gamma), \quad m_i^{(t)} := m_i^{(t-1)} * \gamma + \sum_j^{n_i^{(t)}} E(x)_{i,j}^{(t)}(1 - \gamma), \quad \hat{\mathbf{z}}_{qi}^{(t)} := \frac{m_i^{(t)}}{N_i^{(t)}} \quad (2)$$

where  $sg$  stands for the stop-gradient operation. As per [17, 14], the third loss component in Eq. 1 is replaced by Eq. 2, where  $n_i^{(t)}$  stands for the number of vectors in  $E(\mathbf{x})$  that will be quantized to codebook element  $\hat{\mathbf{z}}_{qi}$ . The hyper-parameters  $\gamma$  and  $\beta$  control the decay of the EMA and the commitment of the encoder output to a certain quantized element respectively.

For the codebook to be perceptually rich, a loss similar to [15, 25] which is formed by the first and second elements of Eq. 1 as defined bellow:

$$\mathcal{L}_{Rec} = \|\mathbf{x} - \hat{\mathbf{x}}\|_1 + \| |FFT(\mathbf{x})| - |FFT(\hat{\mathbf{x}})| \|_2 + \mathcal{L}PIPS_{0.5}(\mathbf{x}, \hat{\mathbf{x}}) \quad (3)$$

Where the first term is a pixel-space L1 norm, the second term is the L2 norm of the image’s Fourier representations based on [26] which aims at facilitating high-frequency feature preservation, the third term is the LPIPS [27] loss using AlexNet applied on 50% of slices on each axis. Lastly, the  $\mathcal{L}_{Adv}$  is based on a Patch-GAN discriminator-based adversarial loss [28, 15], replacing the original loss by the LS-GAN [29] one (see Eq 4):

$$\begin{aligned} \min_D \mathcal{L}_{LSGAN}(D) &= \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [(D(\mathbf{x}) - 1)^2] + \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [(D(G(\hat{\mathbf{x}})))^2] \\ \min_G \mathcal{L}_{LSGAN}(G) &= \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim p_{\mathbf{x}}(\mathbf{z})} [(D(G(\hat{\mathbf{x}})) - 1)^2] \end{aligned} \quad (4)$$

Each of these losses independently contributes to model training stability and reconstruction quality.

### 3.2 Autoregressive Modelling of the Brain

The VQ-VAE model was first trained on T1w MRI images of neurologically healthy subjects from UKB [1] until convergence, and then their  $z_q$  representations were extracted. Afterwards, further fine-tuning on the pathological dataset formed from the baseline T1w MRI scans of ADNI [2] subjects was done until over-fitting was noticed, at which point the ADNI subjects’  $z_q$  representation was also extracted. This paradigm was chosen since in [30] it was shown to either be on par or better compared to training a VQ-VAE model only on the pathological dataset. Furthermore, we aim also to highlight that the pre-trained model can be fine-tuned and learn new morphology, in this case, a pathological one, thus increasing the usefulness of the UKB trained VQ-VAE as a pretrained model for the community.

In order to ensure a higher quality of the Transformer’s samples, the top 1% generated samples, based on the score obtained by averaging the Patch-GAN discriminator output, were used in this work.

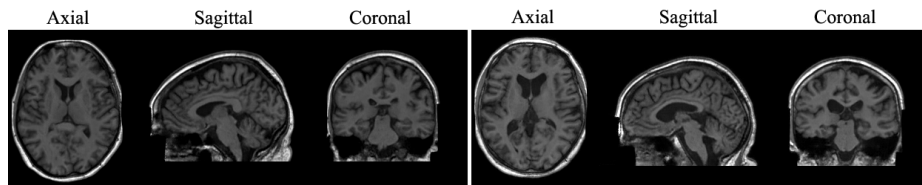
As the VQ-VAE representations cover all phenotypes, a separate Transformer model has been trained on the latent representations of different sub-populations to model individual morphological subgroups. More specifically, to demonstrate the morphological phenotype preservation, the UKB [1] dataset was partitioned into young vs. old sub-populations, and small vs. big ventricles sub-populations. We defined all of these groups based on the first and last of five quantiles based on "age when attended assessment centre" (21003-2.0) and "volume of the ventricular cerebrospinal fluid" (25004-2.0) UKB variables, respectively. To test the preservation of disease morphology, we split the ADNI dataset into cognitively normal (CN) and Alzheimer’s disease (AD) subgroups based on the "diagnosis/scan category assignment field".

## 4 Experiments and Results

The performance of the proposed model is assessed in two ways: first, the quality of generated samples is measured according to image fidelity metrics commonly used in generative models; second, we verify if the morphological characteristics of a population and the differences between sub-populations are preserved when comparing real and synthetic data. We compared our model to a baseline volumetric VAE model. The models by [8, 9, 31, 29] underwent extensive hyperparameter exploration at the original resolutions but failed to converge on our data. Only the VAE results are thus presented as a baseline.

### 4.1 Quantitative Image Fidelity Evaluation

Similarly to [8, 9], we use the FID [12] to assess the visual quality of the generated images. Since originally the metric is based on a pretrained Inception V3 network on 2D natural images, it cannot be applied on 3D volumes directly, so here it is applied on the middle slice of each axis and reported individually. To measure the quality of the 3D samples, batch-wise MMD with a dot product as the kernel is being used as suggested in [8, 9]. Briefly, MMD quantifies the distance between the distributions with finite sample estimates in kernel functions in



**Fig. 1.** Synthetic samples. On the left UKB small ventricles and on the right UKB big ventricles.

the reproducing kernel Hilbert space [13]. Lastly, to estimate the diversity of the generated images, MS-SSIM is being used in a pair-wise fashion between the generated synthetic samples as in [8, 9]. For easy comparison, all metrics have also been calculated between each sub-population’s real images such that a ground truth baseline is also offered.

Across the board, as is described in Table 1, both in regards to the sub-populations and axial, coronal and sagittal slices, the FID of the VQ-VAE model outperforms the VAE baseline by a high margin, showcasing the realistic appearance of the sampled synthetic brains as seen in Fig. 3. The same can be said for the  $bMMD^2$ , where the VQ-VAE is one order of magnitude smaller for UKB sub-populations and substantially better for the ADNI sub-populations. The difference in  $bMMD^2$  performance between VQ-VAE’s UKB and ADNI sub-populations might be because the ADNI dataset is considerably smaller than the UKB one, and to circumvent that, the VQ-VAE compression model was firstly trained on the UKB dataset and then fine-tuned on the ADNI one. Thus, the  $z_q$  representation fed into the Transformer, which is the generative model per se, is not specialised for ADNI, but instead, it tries to encompass it. Finally, MS-SSIM shows that the VQ-VAE achieves a life-like high diversity of samples across all sub-populations, significantly surpassing the VAE. The peculiar case of the ADNI AD sub-population might be attributed to the same cause as the  $bMMD^2$ .

## 4.2 Morphological Evaluation

To evaluate the morphological correctness of the synthetic samples, Voxel-Based Morphometry (VBM) [22] was used to investigate the focal differences in the

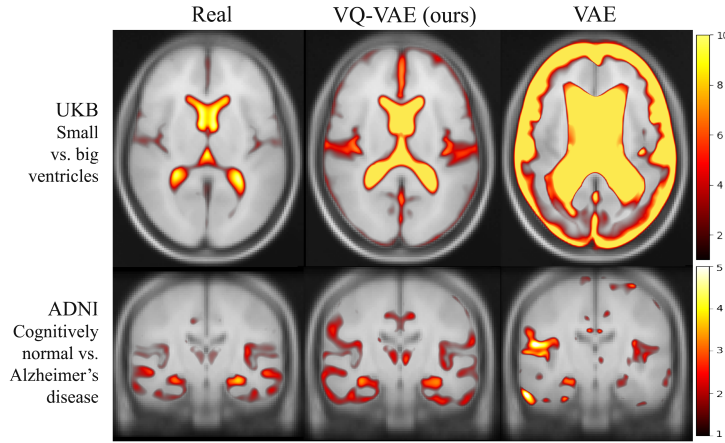
Model	Dataset	Population	FID (Ax Cor Sag)	$bMMD^2$	MS-SSIM
Real	UKB	Young	0.35   0.85   0.42	$0.00208 \pm 0.00026$	$0.65 \pm 0.08$
VQ-VAE (ours)	UKB	Young	31.04   57.19   57.19	$0.00903 \pm 0.00090$	$0.68 \pm 0.03$
VAE	UKB	Young	193.56   302.74   251.34	$0.02757 \pm 0.00091$	$0.15 \pm 0.001$
Real	UKB	Old	1.16   1.42   0.37	$0.00217 \pm 0.00045$	$0.65 \pm 0.07$
VQ-VAE (ours)	UKB	Old	33.68   60.60   78.82	$0.00887 \pm 0.00104$	<b><math>0.67 \pm 0.03</math></b>
VAE	UKB	Old	234.86   289.21   242.18	$0.02622 \pm 0.00044$	$0.15 \pm 0.001$
Real	UKB	Small Ventricles	1.74   1.99   0.87	$0.00220 \pm 0.00044$	$0.67 \pm 0.07$
VQ-VAE (ours)	UKB	Small Ventricles	28.33   58.23   76.68	$0.00892 \pm 0.00106$	<b><math>0.70 \pm 0.04</math></b>
VAE	UKB	Small Ventricles	206.92   318.37   258.17	$0.02836 \pm 0.00078$	$0.14 \pm 0.001$
Real	UKB	Big Ventricles	1.15   1.44   0.53	$0.00231 \pm 0.00041$	$0.64 \pm 0.05$
VQ-VAE (ours)	UKB	Big Ventricles	36.02   57.76   76.51	$0.00937 \pm 0.00069$	$0.68 \pm 0.04$
VAE	UKB	Big Ventricles	215.37   293.97   244.84	$0.02738 \pm 0.00058$	$0.16 \pm 0.001$
Real	ADNI	Cognitively Normal	21.49   17.31   9.34	$0.00123 \pm 0.00021$	$0.56 \pm 0.05$
VQ-VAE (ours)	ADNI	Cognitively Normal	53.88   93.62   112.32	$0.01558 \pm 0.00348$	$0.71 \pm 0.08$
VAE	ADNI	Cognitively Normal	233.59   397.52   421.04	$0.02562 \pm 0.00119$	$0.14 \pm 0.05$
Real	ADNI	Alzheimer’s Diseased	9.08   16.85   13.49	$0.00167 \pm 0.00034$	$0.55 \pm 0.13$
VQ-VAE (ours)	ADNI	Alzheimer’s Diseased	87.75   51.74   90.95	$0.01562 \pm 0.00304$	$0.61 \pm 0.11$
VAE	ADNI	Alzheimer’s Diseased	235.33   332.70   340.78	$0.02804 \pm 0.00177$	$0.12 \pm 0.06$

**Table 1.** The  $bMMD^2$  and MS-SSIM were calculated on 3D generated images while FID was done middle-slices-wise of generated volumes.

brain anatomy of the sub-populations. At the core of VBM stands the application of a generalised linear model and associated statistical tests across all voxels of a group-aligned population, to identify morphological differences in modulated tissue compartment between the selected groups.

The VBM analysis did not factor out any covariates available in the real datasets since the generative process was unconditioned. All t-statistics maps have been corrected to minimise the effects of low variance areas following [32]. As shown in Fig. 2 the t-statistics maps between synthetic images generated by the VQ-VAE strongly agree with the VBM maps of real data, primarily when compared with the VAE baseline. In the UKB small ventricles vs. big ventricles experiment, VQ-VAE again successfully models the ventricular differences correctly compared to the real-data VBM maps, while the VAE model strongly emphasizes them or exacerbates the subarachnoid CSF. Lastly, the morphological differences between cognitively normal vs. AD subjects on the ADNI dataset on the VQVAE generated data strongly preserve the known temporal lobe and hippocampal atrophy patterns associated with AD, producing a VBM t-map that strongly resembles the one from real data. Conversely, the VAE fails to show coherent structural differences in the GM.

Furthermore, we compared the volumes of key brain regions between populations of real and synthetic data. All images were segmented using GIF [23], a robust multi-atlas based probabilistic segmentation model of the human brain which segments the brain into non-overlapping hierarchical 155 regions. Based on probabilistic segmentations, the total volume of each tissue was calculated, and then we ran a two-sided t-test to assess if there was a statistically significant



**Fig. 2.** Thresholded uncorrected VBM t-statistics maps processed as per [32] showcasing the morphological differences between two populations based on real samples, VQ-VAE synthetic samples, and VAE synthetic samples. For UKB small vs. big ventricles modulated CSF tissue segments were used, while for ADNI, cognitively normal vs. AD modulated GM tissue segments were used.

difference between the tissue volumes of the real vs. synthetic populations. The Bonferroni-corrected target p-value was  $2.083\text{e-}05$ .

Model	Dataset	Population	Gray Matter	White Matter	CSF	Deep Gray Matter
Real	UKB	Young	$595_{\pm 32}$	$460_{\pm 29}$	$280_{\pm 21}$	$40_{\pm 3}$
VQ-VAE (ours)	UKB	Young	<b><math>587_{\pm 24}</math></b>	<b><math>472_{\pm 20}</math></b>	<b><math>283_{\pm 11}</math></b>	<b><math>40_{\pm 2}</math></b>
VAE	UKB	Young	$576_{\pm 1}$	$444_{\pm 1}$	$234_{\pm 1}$	$34_{\pm 0}$
Real	UKB	Old	$587_{\pm 31}$	$457_{\pm 29}$	$283_{\pm 22}$	$40_{\pm 3}$
VQ-VAE (ours)	UKB	Old	<b><math>576_{\pm 22}</math></b>	<b><math>465_{\pm 20}</math></b>	$310_{\pm 14}$	<b><math>39_{\pm 2}</math></b>
VAE	UKB	Old	$560_{\pm 1}$	$434_{\pm 1}$	$250_{\pm 1}$	$33_{\pm 0}$
Real	UKB	Small Ventricles	$596_{\pm 30}$	$462_{\pm 27}$	$270_{\pm 17}$	$41_{\pm 3}$
VQ-VAE (ours)	UKB	Small Ventricles	<b><math>594_{\pm 19}</math></b>	$477_{\pm 18}$	$280_{\pm 12}$	<b><math>41_{\pm 2}</math></b>
VAE	UKB	Small Ventricles	$572_{\pm 1}$	$444_{\pm 1}$	$235_{\pm 1}$	$35_{\pm 0}$
Real	UKB	Big Ventricles	$589_{\pm 34}$	$459_{\pm 30}$	$283_{\pm 19}$	$41_{\pm 3}$
VQ-VAE (ours)	UKB	Big Ventricles	$574_{\pm 20}$	<b><math>467_{\pm 17}</math></b>	$307_{\pm 15}$	$39_{\pm 2}$
VAE	UKB	Big Ventricles	$570_{\pm 1}$	$442_{\pm 1}$	$246_{\pm 1}$	$34_{\pm 0}$
Real	ADNI	Cognitively Normal	$530_{\pm 51}$	$430_{\pm 40}$	$309_{\pm 32}$	$40_{\pm 5}$
VQ-VAE (ours)	ADNI	Cognitively Normal	<b><math>554_{\pm 19}</math></b>	<b><math>458_{\pm 18}</math></b>	<b><math>299_{\pm 12}</math></b>	<b><math>39_{\pm 2}</math></b>
VAE	ADNI	Cognitively Normal	<b><math>518_{\pm 5}</math></b>	<b><math>440_{\pm 4}</math></b>	$258_{\pm 3}$	$34_{\pm 1}$
Real	ADNI	Alzheimer's Diseased	$526_{\pm 47}$	$443_{\pm 36}$	$330_{\pm 28}$	$38_{\pm 3}$
VQ-VAE (ours)	ADNI	Alzheimer's Diseased	<b><math>532_{\pm 38}</math></b>	<b><math>446_{\pm 20}</math></b>	$298_{\pm 27}$	<b><math>37_{\pm 3}</math></b>
VAE	ADNI	Alzheimer's Diseased	<b><math>510_{\pm 5}</math></b>	<b><math>443_{\pm 4}</math></b>	$269_{\pm 4}$	$34_{\pm 1}$

**Table 2.** Tissue volumes based on GIF’s probabilistic tissue segmentations. Mean and standard deviations were rounded to the nearest  $10^3$ . The bold values indicate the two-sided t-tests did not pass the statistical significance threshold compared to the real data.

Overall, no significant volume differences were found between real and VQ-VAE samples for most subgroups and tissue types, while significant differences were found for most VAE statistics, demonstrating that the proposed method strongly preserves tissue volumes. The CSF volumes of the VQ-VAE UKB small and big ventricle populations were found to be statistically significantly different from their real counterparts as shown in Table 2, following the VBM results from Fig. 2, and which could explain the increase in the t-statistic observed in the ventricular regions of the synthetic samples. On the other hand, the GM volumes were not statistically significantly different, corroborating the idea that the synthetic t-statistics are closer in magnitude to the real ones. Note that the VAE samples were also found not to be statistically significant in the ADNI AD/CT subset, but this is primarily due to the larger variance and the conservative Bonferroni correction.

## 5 Conclusion

In this work, we propose a scalable and high-resolution volumetric generative model of the brain that preserves morphology. VBM [22] and GIF [23] were used to assess the morphological preservation, while FID [12] and  $bMMD^2$  [13] to measure distribution alignment between synthetic and real samples. We have



shown that the synthetic samples preserve healthy and pathological morphology and that they are realistic images that closely align with the distribution of the real samples. Future work should address the lack of conditioning and the top 1% pruning to increase diversity and provide sampling control. Furthermore, the generative model could be extended for disease progression modelling, disentanglement of style and content, have its privacy preserving capabilities examined, and scaled to include multiple pathologies. To the best of our knowledge, this is the first morphologically preserving generative model of the brain, which paves the way for an unlimited amount of clinically viable data without jeopardizing patient privacy.

**Acknowledgements** WHLP, MG, PB, MJC and PN are supported by Wellcome [WT213038/Z/18/Z]. PTD is supported by the EPSRC Research Council, part of the EPSRC DTP [EP/R513064/1]. FV is supported by Wellcome/EPSRC Centre for Medical Engineering [WT203148/Z/16/Z], Wellcome Flagship Programme [WT213038/Z/18/Z], The London AI Centre for Value-based Healthcare and GE Healthcare. PB is also supported by Wellcome Flagship Programme [WT213038/Z/18/Z] and Wellcome EPSRC CME [WT203148/Z/16/Z]. PN is also supported by the UCLH NIHR Biomedical Research Centre. The models in this work were trained on NVIDIA Cambridge-1, the UK’s largest supercomputer, aimed at accelerating digital biology.

## 6 Appendix

### 6.1 VQ-VAEs

The VQ-VAE model has a similar architecture with [33] but in 3D. The encoder uses strided convolutions with stride 2 and kernel size 4. There are four downsamplings in this VQ-VAE, giving the downsampling factor  $f = 2^4$ . After the downsampling layers, there are three residuals blocks ( $3 \times 3 \times 3$  Conv, ReLU,  $1 \times 1 \times 1$  Conv, ReLU). The decoder mirrors the encoder and uses transposed convolutions with stride 2 and kernel size 4. All convolution layers have 256 kernels. The  $\beta$  in Eq. 1 is 0.25 and the  $\gamma$  in Eq. 2 is 0.5. The codebook size was 2048 while each element’s size was 32.

### 6.2 Transformers

Performer’s<sup>6</sup> [19] has  $L = 24$  layers,  $d = 256$  embedding size, 16 multi-head attention modules (8 are local attention heads with window size of 420), and ReZero gating [34]. Before the raster style ordering input was RAS+ canonical voxel representation oriented.

---

<sup>6</sup> Implementation used: <https://github.com/lucidrains/performer-pytorch>

### 6.3 Losses

VQ-VAE’s pixel-space loss weight is 1.0, perceptual loss’ weight is 0.001, frequency loss’ weight is 1.0. The LPIPS uses AlexNet. Adam has been used as optimizer with an exponential decay of 0.99999. VQ-VAE’s learning rate was 0.000165, discriminator’s learning rate was 0.00005 and Performer’s CrossEntropy learning rate was 0.001.

### 6.4 Datasets

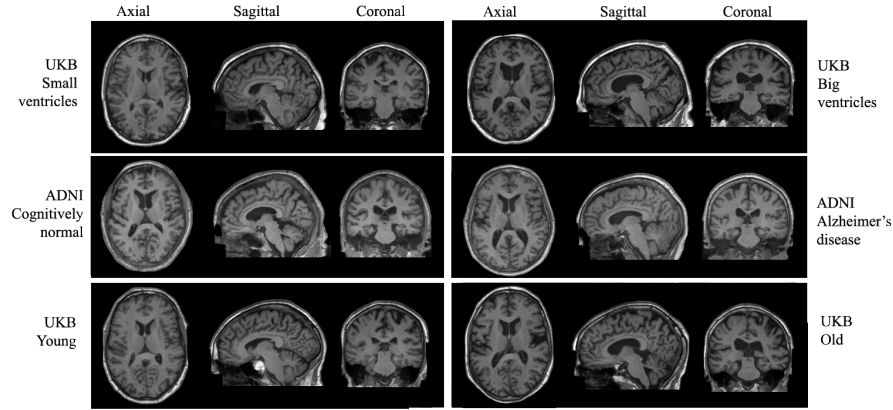
All datasets have been split into training and testing sub-sets. The VQ-VAE UKB sub-sets had 31740 and 3970 subjects respectively, while VQ-VAE ADNI had 648 and 82. All datasets have been first processed with a rigid body registration such that they roughly fit the same field of view. Afterwards, all samples are passed through the following transformations before being fed into the VQ-VAE during training: first, they are being normalized to  $[0,1]$ , then tightly spatially cropped resulting in an image of size (160,224,160), random affine (rotation range 0.04, translation range 2, scale range 0.05), random contrast adjustment (gamma  $[0.99, 1.01]$ ), random intensity shift (offsets  $[0.0,0.05]$ ), random Gaussian noise (mean 0.0, standard deviation 0.02), and finally, the images were thresholded to be in the range  $[0, 1.0]$ . For the Transformer, the UKB and ADNI datasets were split into sub-populations. UKB was split into small ventricles (6388 and 108), big ventricles (6321 and 156), young (6633 and 113), old (5137 and 106), while ADNI was split into cognitively normal (118 and 29) and Alzheimer’s disease (151 and 36). For the Transformer training, each ADNI sample has been augmented 100 times and each augmentation’s index-based representation was used for training it.

### 6.5 VBM analysis

For the Voxel-Based Morphometry (VBM), Statistical Parametric Mapping (SPM) [35] package version 12.7486 was used with MATLAB R2019a. Before running the statistical tests, the images must first undergo unified segmentation where they were spatially normalized to a common template and simultaneously segmented into the Gray Matter (GM), White Matter (WM), and Cerebrospinal fluid (CSF) tissue segments based on prior probability maps and voxel intensities. The unified segmentation was done with the default parameters: Bias Regularisation (light regularisation 0.001), Bias FWHM (60mm cutoff), MRF Parameter (1), Clean Up (Light Clean), Warping Regularisation ( $[0, 0.001, 0.5, 0.05, 0.2]$ ), Affine Regularisation (ICBM space template - European brains), Smoothness (0), Sampling Distance (3). As per standard practice when using VBM, the group-aligned segmentations were modulated to preserve tissue volume, and a smoothing kernel was applied to the modulated tissue compartments to make the data conform to the Gaussian field model that underlines VBM and

to increase the sensitivity to detect structural changes. The smoothing was also done with the default parameters with FWHM ([8, 8, 8]). For the VBM analysis, a Two-sample t-test Design was used, with the following parameters: Independence (Yes), Variance (Unequal), Grand mean scaling (No) and ANCOVA (No). No covariates, masking or global normalisation have been used.

#### Appendix F - Additional Samples



**Fig. 3.** Synthetic samples

## References

- [1] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *PLoS medicine* 12.3 (2015), e1001779.
- [2] Clifford R Jack Jr et al. “The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods”. In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 27.4 (2008), pp. 685–691.
- [3] Amber L Simpson et al. “A large annotated medical image dataset for the development and evaluation of segmentation algorithms”. In: *arXiv preprint arXiv:1902.09063* (2019).
- [4] Chee Keong Chong and Eric Tatt Wei Ho. “Synthesis of 3D MRI brain images with shape and texture generative adversarial deep neural networks”. In: *IEEE Access* 9 (2021), pp. 64747–64760.
- [5] Wanyun Lin et al. “Bidirectional mapping of brain MRI and PET with 3D reversible GAN for the diagnosis of Alzheimer’s disease”. In: *Frontiers in Neuroscience* 15 (2021), p. 357.

- [6] Filip Rusak et al. “3D brain MRI GAN-based synthesis conditioned on partial volume maps”. In: *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer. 2020, pp. 11–20.
- [7] Alice Segato et al. “Data augmentation of 3D brain environment using Deep Convolutional Refined Auto-Encoding Alpha GAN”. In: *IEEE Transactions on Medical Robotics and Bionics* 3.1 (2020), pp. 269–272.
- [8] Gihyun Kwon et al. “Generation of 3D brain MRI using auto-encoding generative adversarial networks”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 118–126.
- [9] Shibo Xing et al. “Cycle Consistent Embedding of 3D Brains with Auto-Encoding Generative Adversarial Networks”. In: *Medical Imaging with Deep Learning*. 2021.
- [10] Li Sun et al. “Hierarchical amortized training for memory-efficient high resolution 3D GAN”. In: *arXiv preprint arXiv:2008.01910* (2020).
- [11] Zhou Wang et al. “Multiscale structural similarity for image quality assessment”. In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.
- [12] Martin Heusel et al. “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In: *Advances in neural information processing systems* 30 (2017).
- [13] Arthur Gretton et al. “A kernel two-sample test”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.
- [14] Ali Razavi et al. “Generating diverse high-fidelity images with vq-vae-2”. In: *Advances in neural information processing systems* 32 (2019).
- [15] Patrick Esser et al. “Taming transformers for high-resolution image synthesis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12873–12883.
- [16] Jiahui Yu et al. “Vector-quantized image modeling with improved vqgan”. In: *arXiv preprint arXiv:2110.04627* (2021).
- [17] Aaron Van Den Oord, Oriol Vinyals, et al. “Neural discrete representation learning”. In: *Advances in neural information processing systems* 30 (2017).
- [18] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [19] Choromanski Krzysztof et al. “Rethinking attention with performers”. In: *Proceedings of ICLR* (2021).
- [20] James Jordon et al. “Synthetic Data—what, why and how?” In: *arXiv preprint arXiv:2205.03257* (2022).
- [21] Cristóbal Esteban et al. “Real-valued (medical) time series generation with recurrent conditional gans”. In: *arXiv preprint arXiv:1706.02633* (2017).
- [22] John Ashburner and Karl J Friston. “Voxel-based morphometry—the methods”. In: *Neuroimage* 11.6 (2000), pp. 805–821.
- [23] M Jorge Cardoso et al. “Geodesic information flows: spatially-variant graphs and their application to segmentation and fusion”. In: *IEEE transactions on medical imaging* 34.9 (2015), pp. 1976–1988.

- [24] Yi Tay et al. “Long Range Arena: A Benchmark for Efficient Transformers”. In: *International Conference on Learning Representations*. 2020.
- [25] Mark S Graham et al. “Transformer-based out-of-distribution detection for clinically safe segmentation”. In: *Medical Imaging with Deep Learning*. 2022.
- [26] Prafulla Dhariwal et al. “Jukebox: A generative model for music”. In: *arXiv preprint arXiv:2005.00341* (2020).
- [27] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.
- [28] Phillip Isola et al. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.
- [29] Xudong Mao et al. “Least squares generative adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2794–2802.
- [30] Petru-Daniel Tudosiu et al. “Neuromorphologically-preserving Volumetric data encoding using VQ-VAE”. In: *arXiv preprint arXiv:2002.05692* (2020).
- [31] Ishaan Gulrajani et al. “Improved training of wasserstein gans”. In: *Advances in neural information processing systems* 30 (2017).
- [32] Gerard R Ridgway et al. “The problem of low variance voxels in statistical parametric mapping; a new hat avoids a ‘haircut’”. In: *Neuroimage* 59.3 (2012), pp. 2131–2141.
- [33] Walter Hugo Lopez Pinaya et al. “Unsupervised Brain Anomaly Detection and Segmentation with Transformers”. In: *Medical Imaging with Deep Learning*. PMLR. 2021, pp. 596–617.
- [34] Thomas Bachlechner et al. “Rezero is all you need: Fast convergence at large depth”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 1352–1361.
- [35] John Ashburner et al. “SPM12 manual”. In: *Wellcome Trust Centre for Neuroimaging, London, UK* 2464 (2014), p. 4.