SILENT FAILURES IN NEUROIMAGING GENERATIVE MODELS: FALSIFICATION-ORIENTED EVALUATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models in neuroimaging are typically evaluated with structural- and intensity-based metrics (SSIM, MSE, PSNR, FID/KID) inherited from natural-image generation and from classical registration, reconstruction, and denoising algorithms whose limited expressive power made such scores a reasonably safe proxy for correctness. We show that this convention becomes actively misleading for deep generative models in neuroimaging, where the scientifically or clinically relevant signal is typically a small deviation from a highly stable, shared anatomical pattern. We replicated two published, openly available models to serve as case studies: BrLP, a latent-diffusion disease progression model, and NT-ViT, an EEG-conditioned fMRI generator, and found that both achieve image-level scores matching their original publications, yet fail to falsify basic claims about task performance once evaluated against trivial baselines (an identity mapping, random subject matching) and task-specific metrics: BrLP's volumetric and directional predictions do not outperform a naive heuristic, and NT-ViT's reconstructions show no measurable relationship to the ground-truth time series despite favorable SSIM and PSNR. Controlled simulations confirm the mechanism: SSIM and MSE remain near ceiling even when all temporal information in an fMRI run is removed, because both are dominated by static, shared anatomy rather than the dynamics of interest. Using a lightweight fMRI tokenizer, we further show that explicit inductive biases in data preprocessing and the training objective expose fidelity/task-performance tradeoffs invisible to any single metric. We argue for evaluation grounded in task-relevant, falsification-oriented metrics, contextualized by task-appropriate baselines, rather than reliance on standard structural and intensity-based scores.

1 INTRODUCTION

The recent successes in image generation algorithms using deep neural networks(Elasri et al., 2022) have inspired a wide range of applications in medical imaging. While many of these algorithms were originally developed for natural images, medical imaging tasks and datasets often present a different set of challenges. This is particularly true in the case of neuroimaging: while widely-employed datasets of natural images (e.g. Imagenet(Deng et al., 2009)) display a large variety of semantically and structurally different content, large datasets of MRI brain imaging are often populated by visually similar samples, in which the clinically relevant variance is found in comparatively small alterations of largely stable anatomical patterns.

Neuroimaging datasets targeting aging, dementia or Alzheimer's disease (e.g. ADNI(Petersen et al., 2010; Jack Jr et al., 2008), OASIS(Marcus et al., 2010)) collect a large number of scans which contain effectively the same anatomical elements, changing in volume based on pathological status, progression, as well as the individual characteristics of the subjects. This is all the more true in the case of functional magnetic resonance imaging (fMRI) data(Friston et al., 1995), which is generally uninformative when analyzing individual 3D scans, where the relevant neurological information is time-dependent and made explicit only after elaborate multi-step preprocessing(Esteban et al., 2019) and analysis pipelines(Zalesky et al., 2010; Zhang et al., 2015; Smitha et al., 2017). Similar considerations extend to other modalities, e.g. diffusion MRI datasets used to study white matter microstructure (the Human Connectome Project(Van Essen et al., 2013), or diffusion protocols within ADNI(Jack Jr et al., 2008)) share a common tract architecture across subjects, with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

pathological or developmental variation expressed as localized shifts in scalar metrics rather than as changes to gross anatomical layout.

Additionally, template-based registration to a common anatomical space (e.g., MNI(Mai & Majtanik, 2017)) is typically assumed as a prerequisite in neuroimaging, from sophisticated fMRI pipelines(Esteban et al., 2019) to structural preprocessing pipelines in animal studies(De Feo & Giove, 2019). The implicit assumption is that alignment to a template is necessary, neutral, and *possible*, which depends entirely on the high structural similarity across brains. Notably, forcing natural image datasets through similar template-based standardization would be immediately recognized as destructive homogenization.

Because of these considerations, we argue that typical image-processing metrics, especially intensity-based or structural metrics, can prove incapable of falsifying a model's fitness for purpose, providing only superficial information about image realism and masking failure modes such as collapse to unconditional generation or inability to beat trivial baselines. We demonstrate these failure modes in two published peer-reviewed models which we take as case studies, BrLP(Puglisi et al., 2024) and NT-ViT(Lanzino et al., 2024). The former is a disease progression model based on latent diffusion, generating the follow-up structural T1 brain scan for healthy, Mild Cognitive Impairment (MCI), and Alzheimer's disease (AD) patients, while the latter is a generative model for fMRI data conditioned on electroencephalographic (EEG) data. In both cases, standard structural metrics fail to detect task- and domain-relevant failure modes: structural similarity can be preserved while temporal dynamics or disease progression patterns are not. We further demonstrate through tokenizer experiments that even simple models can preserve task-specific information when equipped with modality-appropriate inductive biases.

These results raise a general concern about the thoroughness of generative model evaluations in the highly-standardized data distributions found in medical imaging. We propose the following practices to encourage more rigorous, falsification-oriented evaluation and more trustworthy deployment:

- Complement typical computer vision metrics with metrics specific to the targeted machine learning task, and procedures aimed at falsifying direct or implied claims about a model's scope.
- Report falsification against task- and domain-relevant baselines, to contextualize results beyond CV conventions.
- Introduce explicit inductive biases or optimization criteria at the design stage, to proactively reduce the risk of these failure modes.

2 LITERATURE REVIEW

We distinguish "instrumental" generative models, used to support a downstream task such as classification, segmentation, or anomaly detection(Shin et al., 2018; Fernandez et al., 2022; Usman Akbar et al., 2024; Park et al., 2023; Wyatt et al., 2022; Pombo et al., 2023), for which downstream task performance is arguably the only relevant metric, from "terminal" models, where the generated data is itself the intended product of the system, which is our focus in this work.

Beyond instrumental uses, conditioned generative models have been used to predict disease progression(Puglisi et al., 2024; 2025; Yoon et al., 2023; Wan et al., 2026), generate counterfactual images(Dhinagar et al., 2024; Jung et al., 2023), perform general-purpose or condition-specific modality translation(Wolterink et al., 2017; Dar et al., 2019; 2020; Kim et al., 2024; Lanzino et al., 2024; Sargood et al., 2026; Chen et al., 2026; Özbey et al., 2023), and reconstruct accelerated MRI from undersampled k-spaces(Yang et al., 2017; Chung & Ye, 2022; Jalal et al., 2021) or low-resolution inputs(Zhang et al., 2022; Sánchez & Vilaplana, 2018; Xiao et al., 2023).

The surveyed work primarily evaluates image quality using structure- or kernel-based metrics(Huynh & Deshpande, 2024) (SSIM, FID/KID, PSNR) or intensity-based metrics (MSE, RMSE, MAE). While tasks involving clearly visible lesions or large anatomical deformations (e.g. gliomas(Özbey et al., 2023)) may be fairly evaluated via these metrics, we argue that these are especially weak when the variance of interest is expressed as small deviations from a healthy anatomical pattern. A small number of works break away from this pattern and provide stronger evaluation strategies based in neuroimaging: Pombo et al. (2023) utilize Voxel-Based Morphometry as part

of their evaluation strategy, Wan et al. (2026) include the WASABI score(Jafrasteh et al., 2025), developed specifically to evaluate the morphometric plausibility of a distribution of synthetic brain MRIs, Chen et al. (2026) include task-specific evaluations based on the diagnostic accuracy of the generated PET images, and Sargood et al. (2026) include classifier-based verification.

3 EXPERIMENTS

We present results from two complementary case studies demonstrating that standard structural and intensity-based metrics fail to detect whether models preserve clinically or scientifically relevant information. In both BrLP (disease progression modeling) and NT-ViT (EEG-conditioned fMRI generation), visually convincing outputs with good SSIM and MSE scores were incapable of falsifying claims about task performance: disease progression patterns and temporal dynamics, respectively. Conversely, tokenizer experiments demonstrate that introducing task-appropriate inductive biases, such as data standardization and temporal correlation objectives, improves performance on task-relevant metrics without architectural modifications. These results collectively illustrate the necessity of task-aware, falsification-oriented evaluation practices in medical imaging. Full details on datasets, model implementations, and the BSQ tokenizer are provided in Appendix A.

3.1 BRLP: A DISEASE PROGRESSION MODEL IN STRUCTURAL MRI

As a case study on the assessment of terminal structural models we focused on BrLP. BrLP(Puglisi et al., 2024; 2025) is a disease progression model based on latent diffusion, generating a follow-up structural T1 brain scan conditioned on the follow-up time and the condition for healthy, Mild Cognitive Impairment (MCI), and Alzheimer's disease (AD) patients. It was trained on the ADNI(Petersen et al., 2010), OASIS-3(Marcus et al., 2010) and AIBL(Ellis et al., 2009) datasets and published including a code release and open weights, making it an ideal candidate for this case study. BrLP was originally evaluated in terms of MSE and SSIM (image-level) and of the MAE for the volumes of AD-related key regions of interest (ROIs), namely the Hippocampus, Amygdala, Lateral Ventricle, Thalamus and CSF.

We present our results on the ADNI and OASIS-2 datasets, of which only ADNI was originally included in the training data, and utilizing the trained BrLP weights provided by the authors. To provide a baseline for the displayed results we simulated predictions using (1) an identity operation, referring to the trivial baseline of predicting the follow-up to be identical to the initial scan, and two random baselines in which the prediction was replaced by a randomly selected sample of the (2) same age, or (3) the same age and diagnosis. Additionally, for volumetric ROI predictions, we included the predictions of an XGBoost regressor(Chen et al., 2015) trained on volumetric values alone. We evaluated BrLP and compared it with the baselines described above in terms of WASABI scores and directional accuracy of volumetric predictions, in additions to SSIM, MSE, MAE, Fréchet Inception Distance (FID) and Kernel Inception Distance (KID). Directional accuracy was further compared with a naive guess based on the general patterns of ageing and disease trajectory, assuming gray matter regions would shrink and CSF or ventricles increase in volume.

Both random baseline are outperformed by BrLP and the identity baseline, showing considerable evidence of subject-level conditioning. The performance of the identity baseline degrades over larger Δt follow-up intervals (Fig. 1 A, B), and outperforms BrLP according to every metric for $\Delta t \leq 10$ years. Conversely, the performance of BrLP degrades more slowly as Δt increases, remaining closer to the averages of approximately 0.96 SSIM, 0.002 MSE for both datasets (Table 1). For $\Delta t \geq 10$ years, BrLP achieves lower MSE than the identity baseline in 57% of the samples, with an average improvement of 8.1% ($p = 0.0032$, Wilcoxon signed-rank test), however it fails to overtake the identity baseline in terms of SSIM at any Δt. WASABI scores, evaluating the Wasserstein distance between entire distributions of ROI sizes rather than paired predictions, indicate a closer similarity for all baselines compared to BrLP. BrLP exhibited notably elevated FID (45.39 on OASIS-2, 34.66 on ADNI) and KID scores ($27.529 \cdot 10^{-3}$ and $19.355 \cdot 10^{-3}$, respectively), indicating a larger perceptual distance between generated and real brain images compared to the baseline approaches.

Concerning the volumetric predictions in key ROIs, BrLP displays a markedly higher relative MAE compared to the baselines across both datasets and all ROIs (Fig. 1 C, D). Directionally, BrLP underperforms compared to a naive guess in all cases except for the CSF and the Thalamus (Fig. 1, E,

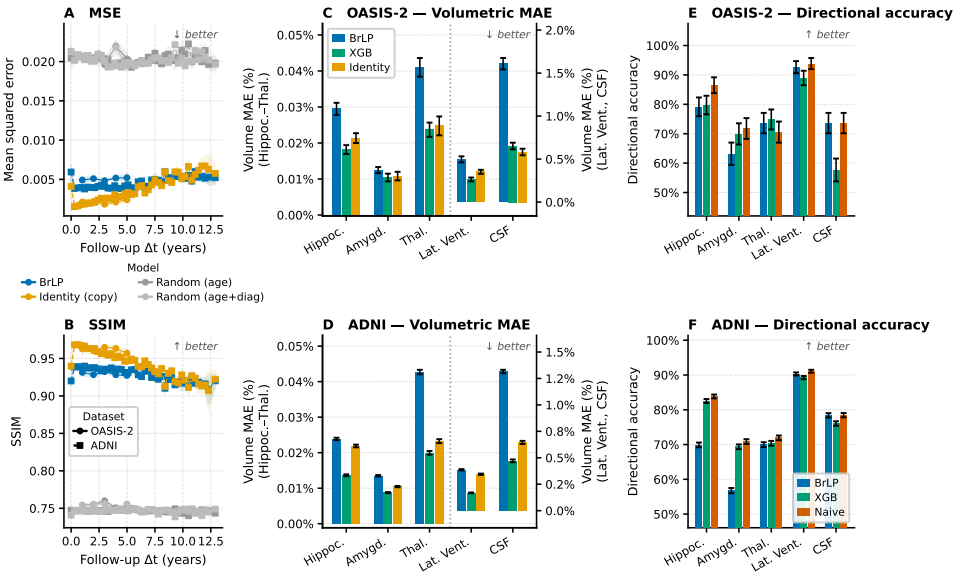

Figure 1: BrLP performance compared to random, identity, and XGB regressor baselines, on the datasets OASIS-2 (external) and ADNI (included in the training distribution). Arrows differentiate between higher-is-better and lower-is-better metrics. A: Mean Squared Error (MSE); B: Structural Similarity Index (SSIM); C, D: relative volumetric Mean Absolute Error (MAE) in the Hippocampus, Amygdala, Thalamus, Lateral Ventricle and Cerebrospinal Fluid (CSF); E, F: directional accuracy, compared to a naive guess (assuming the majority direction of volumetric change). Data was binned in intervals of 0.3 years, bins containing fewer than 3 samples were excluded.

Table 1: Image-level evaluation metrics averaged over all predictions. WASABI is the Wasserstein distance between predicted and ground-truth volumetric distributions of ROI sizes (lower is better except for SSIM, unpaired). Best value per row in **bold**. FID: Fréchet Inception Distance; KID: Kernel Inception Distance

Dataset	Metric	Rand. (age)	Rand. (age+dx)	Identity	BrLP
OASIS-2	MSE	0.0205	0.0205	**0.0021**	0.0050
	SSIM	0.755	0.755	**0.962**	0.930
	WASABI ($\times 10^{-4}$)	1.37	**1.16**	1.34	5.52
	FID	1.34	**1.16**	1.34	45.39
	KID ($\times 10^{-3}$)	0.142	**0.119**	0.167	27.529
ADNI	MSE	0.0204	0.0203	**0.0024**	0.0041
	SSIM	0.747	0.747	**0.959**	0.936
	WASABI ($\times 10^{-4}$)	2.68	**1.06**	1.79	6.51
	FID	0.97	0.56	**0.27**	34.66
	KID ($\times 10^{-3}$)	0.507	0.182	**0.094**	19.355

F). BrLP matches the naive baseline exactly by predicting the CSF volume to increase in nearly every case (100% OASIS-2, 99.6% ADNI) by a a roughly constant magnitude of 6–8%, largely insensitive to the follow-up interval, with only a weak residual dependence on Δt (OASIS-2: +0.60%/year, $R^2 = 0.047$; ADNI: +0.41%/year, $R^2 = 0.089$), suggesting the model has learned a near-fixed directional bias. Conversely in the Thalamus for the OASIS-2 dataset we measured a BrLP 4.42% improvement relative to the naive baseline, however this was not statistically significant (p=0.49, McNemar's test). Comparisons in directional accuracy between the reference XGB regressor and BrLP did not uniformly favor one over the other, however the volumetric predictions of the XGB regressor displayed lower MAE than BrLP and all baselines, with the exception of CSF in OASIS-2, where XGB failed to beat identity. A breakdown of directional accuracy as a function of the follow-up interval does not display any significant deviation from this pattern (see Figure 4 in the Appendix).

3.2 GENERATIVE MODELS FOR FMRI

3.2.1 CHOICE OF METRICS

In addition to standard structural- and intensity-based quality metrics (SSIM, MSE, RMSE), to assess the fidelity of 4D image generation and reconstruction through time we included a set of ROI-based metrics, by comparing the spatially-averaged signal in a set of anatomically-relevant ROIs between the generated images $I*$ and the ground truth I. The ROIs were selected using two commonly-utilized anatomical atlases, the Harvard-Oxford(Desikan et al., 2006) and the Julich(Amunts et al., 2020) atlas. These averaged signals were compared using standard regression metrics (Pearson's correlation, R2 and the Explained Variance (EV)), and the Log-Euclidean distance (LE) between the covariance matrices built from all ROIs from each atlas. The LE distance(Arsigny et al., 2006) is defined as the Frobenius norm of the difference of the matrix logarithms, a metric used to measure the similarity between symmetric positive definite matrices, often used to compare covariance matrices(Jayasumana et al., 2015; Ilea et al., 2018; Li et al., 2013; Wang et al., 2012) and utilized here to evaluate the preservation of functional connectivity given a set of ROIs:

$\mathrm{LE}(I^*, I) = ||\log(\mathrm{cov}(I^*)) - \log(\mathrm{cov}(I))||_F$ where F refers to the Frobenius norm and cov to the covariance matrix of the timeseries obtained via the two anatomical atlases. Following Calhas & Henriques (2020), we further included the cosine-distance-based Cosine Flattened Voxels (CFV) and Log-Cosine Flattened Voxels (LCFV) as global metrics evaluating the time-dependent dynamics independently of the choice of ROIs:

$$\mathrm{CFV}(I^*, I) = \cos(\text{flatten}(I^*), \text{flatten}(I))$$
$$\mathrm{LCFV}(I^*, I) = \log(1 - \mathrm{CFV}(I^*, I))$$

3.2.2 SIMULATED UNCONDITIONAL GENERATION

To evaluate metric sensitivity to the removal of time-dependent dynamics independently of structural content, we simulated reconstruction of fMRI sample I_j in three ways: adding Gaussian noise to the original data ($I_j^* = I_j + N(0, \sigma^2)$, Data + Noise, DN), to time-averaged data ($I_j^* = \bar{I}_j + N(0, \sigma^2)$, Mean + Noise, MN), or linearly mixing with a random sample k ($I_j^* = (1 - \mu)I_j + \mu I_k$, data replacement, DR). The standard deviation σ varied from 10^{-2} to 10^2, and mixing factor μ from 0 to 1, with $\mu = 1$ corresponding to complete replacement. Samples were standardized to z-scores for meaningful noise-data comparison, with values averaged across 60 replicas. For this task we utilized data from the Human Connectome Project dataset (HCP)(Van Essen et al., 2013), S1200 release, because of its established influence in fMRI literature.

At the low-noise regimen, only a select number of metric were effectively able to differentiate between DN and MN cases. SSIM was insensitive to the removal of all time-dependent dynamics, yielding close to ideal results (1) for MN experiments (Fig. 2). Similarly MSE was weakly sensitive to the removal of temporal information, detecting differences no larger than $10^{-1} \sim 10^{-2}$ at $\sigma = 10^{-1}$. While CFV behaved similarly to SSIM, LCFV could differentiate between DN and MN at low noise (lower-is-better metric, approximately -11 *vs.* -4).

In contrast R^2 and EV were highly sensitive to time-dependent information, differentiated mainly by random-mixing response: R^2 degraded strongly for low μ, EV decreased gradually. Both degraded rapidly for $\sigma \geq 0.1$. Pearson correlation and LE were highly sensitive: correlation dropped to zero while LE remained between 10 and 10^2. Both metrics smoothly transitioned from high quality ($\sigma \sim 10^{-2}$) to noise-indistinguishable ($\sigma \sim 10^2$).

In the high-noise regimen, all metrics except Pearson, LCFV, EV and R^2 distinguished high-noise from DR, detecting greater similarity to unrelated realistic data than noise-dominated reconstructions, while DN and MN converged to the same values. Pearson remained the most robust to noise, registering a small but detectable difference between DN and MN up to $\sigma \sim 5$, while most metrics cease to distinguish MN from DN between $\sigma = 0.1$ and $\sigma = 1$. While SSIM was sensitive to the addition of noise in general, degrading to ≈ 0.2 equally in both MN and DN, it was largely insensitive to DR, displaying $\mathrm{SSIM} \approx 0.8$ at full replacement.

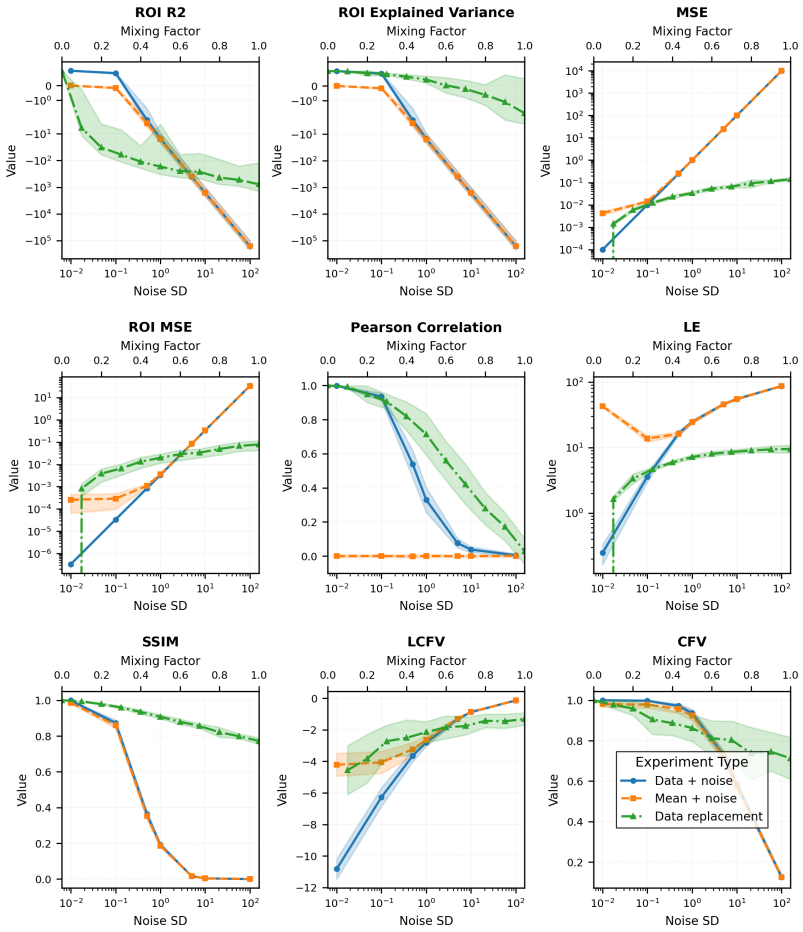

Figure 2: Quality metrics for simulated reconstruction experiments: original plus Gaussian noise (blue), temporal mean plus Gaussian noise (orange), or linear mixing with a random volume from the same dataset (green). Bottom x-axis: noise standard deviation σ (log scale). Top x-axis: mixing factor μ (0=original, 1=replacement). The ROI prefix indicates metrics computed on anatomically-averaged timeseries.

3.2.3 Case study: conditional image generation

To evaluate the fidelity of the reconstruction of time-dependent dynamics in fMRI image generation, we replicated NT-ViT(Lanzino et al., 2024), an EEG-conditioned generative neural network for fMRI data, using the code provided by the authors. Consistently with its original implementation, NT-ViT was trained on the EEG-fMRI Oddball(Walz et al., 2014; 2013; Conroy et al., 2013; Walz et al., 2015) and NODDI(Deligianni et al., 2016; 2014) datasets, and evaluated using a Leave-One-Out (LOO) scheme.

Evaluation results are displayed in Table 2. The measured SSIM of 0.683 and 0.626 (respectively in the Oddball and NODDI datasets) were comparable with the ones originally published (0.627 and 0.581), and likewise we measured RMSE of 0.078 and 0.094 *vs.* the original 0.07 and 0.08, and PSNR of 22.20, 20.63 (*vs.* 23.33, 21.56). While structural and intensity-based metrics could suggest a good visual matching with the ground truth data, regression metrics for the ROI averages (R2, EV, Pearson's correlation) as well as the LE distance and LCFV do not appear to indicate the reconstruction of any time-dependent dynamics. We found, across all atlases and datasets, $\mathrm{R2} \leq -78$, $\mathrm{EV} \leq -0.4$, $\mathrm{LE} \geq 25$, $|\rho| \leq 0.0209$, and $\mathrm{LCFV} \approx -0.5$. Likewise, NT-ViT demonstrated substantially degraded reconstruction quality according to perceptual metrics, with FID scores of 109.58 on Oddball and 139.84 on NODDI, alongside correspondingly elevated KID values (0.0679 and 0.1106, respectively).

Table 2: NT-ViT reconstruction quality on the Oddball and NODDI datasets. Image-level metrics are computed for the entire brain; ROI-level metrics are averaged within each atlas region and shown per atlas. Kernel-based metrics FID, KID were calculated per-slice.

Metric	Oddball		NODDI	
SSIM	0.6833		0.6261	
RMSE	0.0782		0.0943	
PSNR	22.2042		20.6374	
MSE	0.0062		0.0092	
LCFV	-0.0522		-0.0568	
CFV	0.9492		0.9450	
FID	109.5775		139.8381	
KID	0.0679		0.1106	
	Harvard-Oxford	Julich	Harvard-Oxford	Julich
Pearson Correlation	0.0004	0.0051	0.0209	-0.0120
ROI R2	-85.4412	-78.2320	-115.5166	-105.0916
ROI EV	-7.2251	-7.6601	-1.1121	-0.4734
ROI MSE	0.0041	0.0058	0.0068	0.0068
LE	25.1980	46.0322	47.5655	77.2173

3.2.4 IMAGE TOKENIZER

To experiment with optimization objectives, data standardization strategies and evaluation metrics in a controlled setting, under a simple reconstruction objective, we trained a lightweight BSQ-based(Zhao et al., 2024) convolutional tokenizer as a pure encoder-decoder and evaluated it according to the fidelity metrics discussed in section 3.2.1. The fidelity metrics in this case would represent the fidelity ceiling for an image generation model relying on that tokenizer, e.g. in the style of MaskGIT(Chang et al., 2022). After training the model on the HCP dataset, we additionally evaluated it by training a logistic regressor classifier on the connectivity matrices of the reconstructed 4D images, to identify the HCP behavioral task (emotion, gambling, language, motor, relational, rest, social, working memory).

To introduce an inductive bias or objective aligned with the preservation of time-dependent fMRI dynamics, we trained different variants of the models differentiated by the stride of the first downsampling 3D convolution (6 or 12 voxels in both receptive field and stride, affecting the coarseness of the tokenizer), the scaling strategy (MinMax or difference from the mean volume of the 4D run, here referred to as Diff), and the use of a loss function introducing Pearson's correlation directly as an optimization target (SPC loss). More details on the tokenizer are presented in section A.3.

Reconstruction fidelity metrics and classification performance revealed marked divergence across variants (Fig. 3) and exposed considerable trade-offs. MinMax scaling consistently outperformed Diff scaling on classification accuracy, with accuracy approximately 0.93-0.96 compared to 0.81-0.86 for Diff, while Diff-scaled reconstructions matched the original-data baseline across both tokenizer resolutions. Introducing SPC loss in the presence of MixMax scaling improved Pearson correlation, at the cost of modestly reducing classification accuracy and increasing LE distance. Without SPC loss, Diff scaling matched or exceeded MinMax on SSIM and MSE, though with greater variance, and achieved better Pearson correlation and LE distance. convolution stride exhibited minimal effect on classification performance, while otherwise degrading both structural and time-sensitive metrics. The bimodal distributions visible in the violin plots reflect measurements computed across the two different brain atlases.

4 DISCUSSION

In our work we hypothesized that the high degree of standardization and anatomical similarity in neuroimaging datasets would result in intensity- or structure-based metrics being unable to falsify claims about task performance when coarse anatomy can act as a confounder. We tested this hypothesis in a number of experiments and case studies, replicating two published models (BrLP and NT-ViT) utilizing the code and, when available, the model weights provided by the authors.

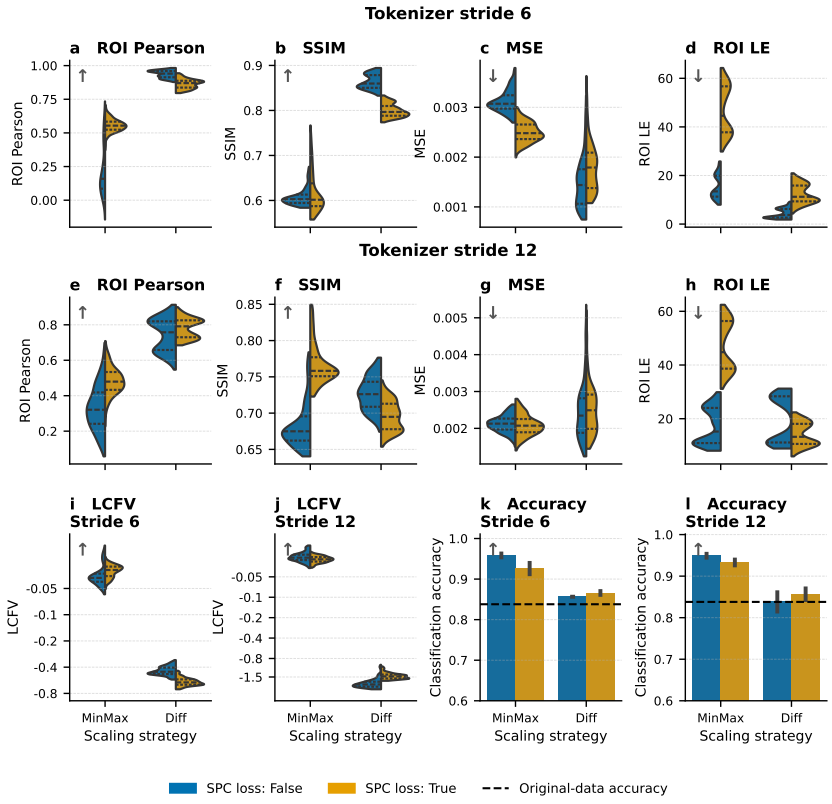

Figure 3: Image reconstruction quality for fMRI tokenizers by stride, quality metric, use of the SPC loss, and data scaling strategy. ROI indicates metrics calculated based on the localized averages in a selection of ROIs defined by the Harvard-Oxford and the Julich atlases; Accuracy refers to the accuracy of a logistic regressor classifier trained to identify the fMRI task in each run, based on the functional connectivity of the reconstructed images. Arrows differentiate between higher-is-better and lower-is-better metrics.

The central issue raised by our results is not whether the evaluated models produced poor results or represented unpromising directions, but pertains to the metrics used to support claims about task performance. BrLP's image-level scores (SSIM ≈ 0.96, MSE ≈ 0.002, Table 1) and NT-ViT's structural and intensity-based scores (Table 2) are, by conventional computer vision standards, indicative of successful reconstruction. Yet, metrics and baselines tailored to the quantity each model was built to recover told a different story, with BrLP's volumetric and directional predictions failing to consistently improve on a naive guess or an identity mapping (Fig. 1), and NT-ViT's reconstructions showing no measurable relationship to the ground-truth time series ($R^2 \leq -78$, $|\rho| \leq 0.0209$).

Shared anatomy dominates the signal for structural metrics. The simulated unconditional generation experiments (Fig. 2) offer a direct, controlled explanation for this pattern. When all time-dependent information in an fMRI run is replaced by its temporal mean plus noise, SSIM remains close to its ceiling and MSE is only weakly perturbed at low noise levels, indicating that both metrics are dominated by the static anatomical content that is largely shared across samples; the same insensitivity appears, in a milder form, under full replacement with an unrelated real sample (SSIM ≈ 0.8 at $\mu = 1$). Metrics that are instead sensitive to the temporal structure specifically, such as Pearson's correlation, the Log-Euclidean distance between connectivity matrices, R^2, and explained variance, tracked the manipulation closely and only degraded once dynamics were actually removed or replaced. The same dissociation is visible in the two case studies: metrics dominated by the shared anatomical structure appeared favorable, while those that revealed the failure were the ones designed to isolate the task-relevant component of the signal. More broadly, this reliance on structural and intensity-based metrics reflects an assumption that used to hold for algorithms applied

in neuroimaging, but does not carry over to deep generative models: classical registration (Avants et al., 2009; Jenkinson et al., 2012; Jenkinson & Smith, 2001; Andersson et al., 2007; Ashburner & Friston, 2005) is constrained to smooth spatial deformations, parallel-imaging and compressed-sensing reconstruction (Lustig et al., 2008) is constrained by consistency with the acquired k-space data, and denoising or bias-field correction (Veraart et al., 2016; Tustison et al., 2010) rely on explicit physical or statistical models of the corruption being removed. In each case the algorithm's narrow expressive power made a high intensity- or structure-based score a reasonably safe proxy for correctness, a guarantee unconstrained deep generative models do not offer.

Conditioning and evaluation limits. For BrLP we detected clear evidence of conditioning to the anatomical prior: both random baselines were consistently outperformed. However, once a trivial identity-mapping comparison is introduced, none of the metrics originally used to characterize the model are able to establish that its predictions provide more information than what could be obtained from the baseline scan and a simple heuristic, with the identity baseline achieving performance competitive with BrLP on every reported metric. The additional value provided by a model of the complexity of a deep generative network is not evident in terms of volumetric estimates and directional accuracy of disease development when compared to either a regressor trained exclusively on volumetric information, or an identity mapping coupled with a guess based on the general patterns of the disease trajectory, according to the same evaluation protocol used to report BrLP as successful.

Implications for deployment in clinical trials. Disease progression models have been proposed as tools for prognostic enrichment and clinical trial cohort simulation (Puglisi et al., 2025); we are not asserting that BrLP is unsuitable for every possible use, nor questioning the plausibility of what could be a promising modeling approach. However, our results suggest that evaluations based on structural- and intensity-based metrics alone would likely be overconfident about whether such a model outperforms a cost-free heuristic based on the typical disease trajectory, with serious implications where the intended use is the prediction of an individualized disease trajectory.

Inductive biases and tradeoffs. Introducing an inductive bias at the level of data preprocessing or the optimization objective, and reporting a battery of metrics rather than a single score, is itself a way of exposing tradeoffs that would otherwise go unnoticed: in the tokenizer case study (Fig. 3), no single scaling strategy or training objective dominated jointly on fidelity and task-relevant metrics. Metrics chosen to target domain-relevant temporal structure and downstream measures, such as classifier accuracy, do not necessarily move together, and improving one may come at the expense of another. Both FID and KID tracked the failures identified through other means, elevated for BrLP and NT-ViT alike, suggesting they can be a useful complement to SSIM and MSE. Nonetheless, their application to medical images implies several compromises. The feature extractors were trained on natural rather than medical images, and further require the 2D-slice decomposition of volumetric or 4D data. Their relevance beyond assessing broad perceptual realism has been questioned, and the development of kernel-based scores tailored to medical generative models remains at a comparatively early stage (Woodland et al., 2024). The practical implication is that evaluation batteries need to be chosen deliberately, tailored to the intended purpose of the model, and with baselines that contextualize the observed values rather than comparing models only to one another.

Limitations. We restricted our experiments to two published, reproducible models spanning structural and functional neuroimaging. Both provide open code, and BrLP additionally distributes trained weights. Doing so grounded every claim in a concrete, replicable comparison rather than a general assertion; this is a strength for reproducibility, but means our directly re-evaluated models are drawn from the more transparent end of the literature we surveyed. We do not read these results as a failure specific to BrLP or NT-ViT. The underlying property we identify, that a dataset's dominant source of image variance is a stable anatomical pattern while clinically relevant variance is a comparatively small deviation from it, is not particular to T1-weighted structural scans or EEG-fMRI synthesis, as discussed in the Introduction. Finally, the simulated unconditional-generation experiments (Fig. 2) are a controlled simplification isolating each metric's sensitivity to temporal information, real model failures may combine loss of temporal dynamics with other distortions not captured by this simulation. The XGBoost regressor comparison is diagnostic and introduced to contextualize the results, and does not represent a proposed alternative solution.

AI USE STATEMENT

In this work, we used generative AI tools for accelerating the writing of glue code and evaluation code in the reimplementation of BrLP and NT-ViT, as well as the dataset classes used for our tokenizer experiments. The code as well as its generated outputs were manually reviewed by the authors, and we verified that the quality measures obtained were compatible with the ones presented by the authors in their original work. Generative AI was further use to polish the manuscript and the presentation of the released Jupyter notebooks concerning the BrLP results.

REPRODUCIBILITY STATEMENT

The experimental code utilized in this work is available at https://anonymous.4open.science/r/failure-modes-8757 (anonymized for the review process). The datasets utilized in this work are available from their respective authors (see section A.1).

REFERENCES

Katrin Amunts, Hartmut Mohlberg, Sebastian Bludau, and Karl Zilles. Julich-brain: A 3d probabilistic atlas of the human brain's cytoarchitecture. *Science*, 369(6506):988–992, 2020.

Jesper LR Andersson, Mark Jenkinson, Stephen Smith, et al. Non-linear registration aka spatial normalisation fmrib technial report tr07ja2. *FMRIB Analysis Group of the University of Oxford*, 2007.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 56(2):411–421, 2006.

John Ashburner and Karl J Friston. Unified segmentation. *Neuroimage*, 26(3):839–851, 2005.

Brian B Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ants). *Insight j*, 2: 1–35, 2009.

David Calhas and Rui Henriques. Eeg to fmri synthesis: Is deep learning a candidate? *arXiv preprint arXiv:2009.14133*, 2020.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11305–11315. IEEE, 2022.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

Zhigeng Chen, Sheng Bi, Yi Shan, Feng Wang, Yong Wang, Zhongyuan Qi, Tao Wang, Xiaoyuan Li, Shengnan Li, Huanhui Xiao, et al. Mri-to-pet synthesis via deep learning for amyloid-β quantification in alzheimer's disease. *European Radiology*, pp. 1–13, 2026.

Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80:102479, 2022.

Bryan R Conroy, Jennifer M Walz, and Paul Sajda. Fast bootstrapping and permutation testing for assessing reproducibility and interpretability of multivariate fmri decoding models. *PloS one*, 8 (11):e79271, 2013.

Salman UH Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Cukur. Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE transactions on medical imaging*, 38(10):2375–2388, 2019.

Salman UH Dar, Mahmut Yurt, Mohammad Shahdloo, Muhammed Emrullah Ildız, Berk Tınaz, and Tolga Çukur. Prior-guided image reconstruction for accelerated multi-contrast mri via generative adversarial networks. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1072–1087, 2020.

Riccardo De Feo and Federico Giove. Towards an efficient segmentation of small rodents brain: a short critical review. *Journal of neuroscience methods*, 323:82–89, 2019.

Fani Deligianni, Maria Centeno, David W Carmichael, and Jonathan D Clayden. Relating resting-state fmri and eeg whole-brain connectomes across frequency bands. *Frontiers in neuroscience*, 8:258, 2014.

Fani Deligianni, David W Carmichael, Gary H Zhang, Chris A Clark, and Jonathan D Clayden. Noddi and tensor-based microstructural indices as predictors of functional connectivity. *Plos one*, 11(4):e0153404, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

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

Mohamed Elasri, Omar Elharrouss, Somaya Al-Maadeed, and Hamid Tairi. Image generation: A review. *Neural Processing Letters*, 54(5):4609–4646, 2022.

Kathryn A Ellis, Ashley I Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T Lautenschlager, Nat Lenzo, Ralph N Martins, Paul Maruff, et al. The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. *International psychogeriatrics*, 21(4):672–687, 2009.

Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

Virginia Fernandez, Walter Hugo Lopez Pinaya, Pedro Borges, Petru-Daniel Tudosiu, Mark S Graham, Tom Vercauteren, and M Jorge Cardoso. Can segmentation models be trained with fully synthetically generated data? In *International workshop on simulation and synthesis in medical imaging*, pp. 79–90. Springer, 2022.

Karl J Friston, Andrew P Holmes, JB Poline, PJ Grasby, SCR Williams, Richard SJ Frackowiak, and Robert Turner. Analysis of fmri time-series revisited. *Neuroimage*, 2(1):45–53, 1995.

Nguyen Huynh and Gopikrishna Deshpande. A review of the applications of generative adversarial networks to structural and functional mri based diagnostic classification of brain disorders. *Frontiers in Neuroscience*, 18:1333712, 2024.

Ioana Ilea, Lionel Bombrun, Salem Said, and Yannick Berthoumieu. Covariance matrices encoding based on the log-euclidean and affine invariant riemannian metrics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 393–402, 2018.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

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

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. *Advances in neural information processing systems*, 34:14938–14954, 2021.

Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtash Harandi. Kernel methods on riemannian manifolds with gaussian rbf kernels. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2464–2477, 2015.

Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.

Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

Euijin Jung, Miguel Luna, and Sang Hyun Park. Conditional gan with 3d discriminator for mri generation of alzheimer's disease progression. *Pattern Recognition*, 133:109061, 2023.

Sooyoung Kim, Joonwoo Kwon, Junbeom Kwon, Jungyoun Janice Min, Sangyoon Bae, Yuewei Lin, Shinjae Yoo, and Jiook Cha. Macro2micro: A rapid and precise cross-modal magnetic resonance imaging synthesis using multi-scale structural brain similarity. *arXiv preprint arXiv:2412.11277*, 2024.

Romeo Lanzino, Federico Fontana, Luigi Cinque, Francesco Scarcello, and Atsuto Maki. Neural transcoding vision transformers for eeg-to-fmri synthesis. In *European Conference on Computer Vision*, pp. 53–70. Springer, 2024.

Peihua Li, Qilong Wang, Wangmeng Zuo, and Lei Zhang. Log-euclidean kernels for sparse representation and dictionary learning. In *Proceedings of the IEEE international conference on computer vision*, pp. 1601–1608, 2013.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72–82, 2008.

Jürgen K Mai and Milan Majtanik. *Human brain in standard MNI space: a comprehensive pocket atlas*. Elsevier, 2017.

Daniel S Marcus, Anthony F Fotenos, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *Journal of cognitive neuroscience*, 22(12):2677–2684, 2010.

Steen Moeller, Essa Yacoub, Cheryl A Olman, Edward Auerbach, John Strupp, Noam Harel, and Kâmil Uğurbil. Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. *Magnetic resonance in medicine*, 63(5):1144–1153, 2010.

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Cukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023.

Seungju Park, Kyung Hwa Lee, Beomseok Ko, and Namkug Kim. Unsupervised anomaly detection with generative adversarial networks in mammography. *Scientific Reports*, 13(1):2925, 2023.

Ronald Carl Petersen, Paul S Aisen, Laurel A Beckett, Michael C Donohue, Anthony Collins Gamst, Danielle J Harvey, Clifford R Jack Jr, William J Jagust, Leslie M Shaw, Arthur W Toga, et al. Alzheimer's disease neuroimaging initiative (adni) clinical characterization. *Neurology*, 74(3): 201–209, 2010.

Guilherme Pombo, Robert Gray, M Jorge Cardoso, Sebastien Ourselin, Geraint Rees, John Ashburner, and Parashkev Nachev. Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models. *Medical Image Analysis*, 84: 102723, 2023.

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

Lemuel Puglisi, Daniel C Alexander, Daniele Ravì, Alzheimer's Disease Neuroimaging Initiative, et al. Brain latent progression: Individual-based spatiotemporal disease progression on 3d brain mris via latent diffusion. *Medical image analysis*, pp. 103734, 2025.

Irina Sánchez and Verónica Vilaplana. Brain mri super-resolution using 3d generative adversarial networks. *arXiv preprint arXiv:1812.11440*, 2018.

Alec Sargood, Lemuel Puglisi, James H Cole, Neil P Oxtoby, Daniele Ravì, and Daniel C Alexander. Cocolit: Controlnet-conditioned latent image translation for mri to amyloid pet synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8778–8786, 2026.

Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *International workshop on simulation and synthesis in medical imaging*, pp. 1–11. Springer, 2018.

Karavallil A Smitha, K Akhil Raja, KM Arun, PG Rajesh, Bejoy Thomas, TR Kapilamoorthy, and Chandrasekharan Kesavadas. Resting state fmri: A review on methods in resting state connectivity analysis and resting state networks. *The neuroradiology journal*, 30(4):305–317, 2017.

Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

Muhammad Usman Akbar, Måns Larsson, Ida Blystad, and Anders Eklund. Brain tumor segmentation using synthetic mr images-a comparison of gans and diffusion models. *Scientific Data*, 11 (1):259, 2024.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Jelle Veraart, Els Fieremans, and Dmitry S Novikov. Diffusion mri noise mapping using random matrix theory. *Magnetic resonance in medicine*, 76(5):1582–1593, 2016.

Jennifer M Walz, Robin I Goldman, Michael Carapezza, Jordan Muraskin, Truman R Brown, and Paul Sajda. Simultaneous eeg-fmri reveals temporal evolution of coupling between supramodal cortical attention networks and the brainstem. *The Journal of Neuroscience*, 33(49):19212–19222, 2013.

Jennifer M Walz, Robin I Goldman, Michael Carapezza, Jordan Muraskin, Truman R Brown, and Paul Sajda. Simultaneous eeg–fmri reveals a temporal cascade of task-related and default-mode activations during a simple target detection task. *Neuroimage*, 102:229–239, 2014.

Jennifer M Walz, Robin I Goldman, Michael Carapezza, Jordan Muraskin, Truman R Brown, and Paul Sajda. Prestimulus eeg alpha oscillations modulate task-related fmri bold responses to auditory stimuli. *NeuroImage*, 113:153–163, 2015.

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

Ruiping Wang, Huimin Guo, Larry S Davis, and Qionghai Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2496–2503. IEEE, 2012.

Jelmer M Wolterink, Anna M Dinkla, Mark HF Savenije, Peter R Seevinck, Cornelis AT Van Den Berg, and Ivana Išgum. Deep mr to ct synthesis using unpaired data. In *International workshop on simulation and synthesis in medical imaging*, pp. 14–23. Springer, 2017.

McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P Yung, Ankit B Patel, et al. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 87–97. Springer, 2024.

Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 650–656, 2022.

Yueyue Xiao, Chunxiao Chen, Liang Wang, Jie Yu, Xue Fu, Yuan Zou, Zhe Lin, and Kunpeng Wang. A novel hybrid generative adversarial network for ct and mri super-resolution reconstruction. *Physics in Medicine & Biology*, 68(13):135007, 2023.

Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, et al. Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2017.

Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. Sadm: Sequence-aware diffusion model for longitudinal medical image generation. In *International Conference on Information Processing in Medical Imaging*, pp. 388–400. Springer, 2023.

Andrew Zalesky, Alex Fornito, and Edward T Bullmore. Network-based statistic: identifying differences in brain networks. *Neuroimage*, 53(4):1197–1207, 2010.

Kuan Zhang, Haoji Hu, Kenneth Philbrick, Gian Marco Conte, Joseph D Sobek, Pouria Rouzrokh, and Bradley J Erickson. Soup-gan: Super-resolution mri using generative adversarial networks. *Tomography*, 8(2):905–919, 2022.

Linlin Zhang, Michele Guindani, and Marina Vannucci. Bayesian models for functional magnetic resonance imaging data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7 (1):21–41, 2015.

Yue Zhao, Yuanjun Xiong, and Philipp Krähenbühl. Image and video tokenization with binary spherical quantization. *arXiv preprint arXiv:2406.07548*, 2024.

A Implementation details

A.1 Datasets

A.1.1 fMRI

HCP We utilized data from the WU-Minn Human Connectome Project dataset, S1200 release(Van Essen et al., 2013). We included only 3T fMRI data acquired using a gradient-echo Echo-Planar Imaging sequence (EPI)(Moeller et al., 2010) with a TR of $720\,\mathrm{ms}$, using the preprocessed scans aligned in standard MNI space(Mai & Majtanik, 2017), at a $2\,\mathrm{mm}$ isotropic resolution, and including 18 runs for each subject. These runs included seven task runs (emotion processing, relational processing, social cognition, language, motor, working memory and gambling tasks) and two resting state runs. These nine samples were repeated with right-to-left and left-to-right phase encoding directions, for a total of 18 4D fMRI volumes per subject.

NODDI This dataset comprises simultaneous resting-state EEG-fMRI data from 16 adult subjects (mean age 32.8 years) acquired during eyes-open fixation on a 1.5 T scanner, along with diffusion-weighted imaging data optimized for the NODDI model(Deligianni et al., 2016; 2014). fMRI data consisted of 300 volumes acquired with TR/TE 2160/30 ms at $3.3 \times 3.3 \times 4.0$ mm resolution. Concurrent EEG was recorded with a 64-channel cap at 1000 Hz sampling rate.

Oddball This dataset comprises simultaneous EEG-fMRI data from 17 subjects performing auditory and visual oddball tasks(Walz et al., 2014; 2013; Conroy et al., 2013; Walz et al., 2015). fMRI data consisted of three runs per task acquired on a 3 T scanner using a single-channel coil with EPI acquisition (170 TRs per run, TR 2 s, 32 slices, 3 × 3 × 4 mm resolution). EEG was recorded simultaneously using a custom MR-compatible system with 49 channels at 1000 Hz sampling rate, synchronized with scanner pulses.

A.1.2 STRUCTURAL MRI

ADNI We utilized T1-weighted structural MRI scans from the Alzheimer's Disease Neuroimaging Initiative, a longitudinal study with baseline and follow-up visits(Jack Jr et al., 2008; Petersen et al., 2010). Data included subjects from three diagnostic groups: cognitively normal, Mild Cognitive Impairment (MCI), and Alzheimer's disease (AD). All scans were acquired at 1.5 T or 3 T on standardized protocols and preprocessed to MNI template space, enabling disease progression modeling across the follow-up period.

OASIS-2 We utilized T1-weighted structural MRI scans from the Open Access Series of Imaging Studies (OASIS-2), a longitudinal collection of 150 subjects aged 60–96 with multiple follow-up visits separated by at least one year(Marcus et al., 2010). The dataset includes 373 imaging sessions with 72 cognitively normal subjects throughout the study, 64 subjects with dementia at baseline (including 51 with mild-to-moderate Alzheimer's disease), and 14 subjects who progressed from cognitively normal to demented status during follow-up. All scans are 1 × 1 × 1 mm resolution T1-weighted images acquired on a 1.5 T scanner.

A.2 BRLP

Brain Latent Progression (BrLP) is a spatiotemporal disease progression model combining latent diffusion with domain knowledge of neurodegeneration. The architecture comprises four components: i) a Latent Diffusion Model (LDM) that learns the distribution of brain MRIs conditioned on covariates (age, sex, cognitive status); ii) a ControlNet that conditions the LDM on subject-specific brain anatomy by incorporating latent representations from baseline MRIs; iii) an auxiliary model that predicts volumetric changes in AD-related brain regions over time (using either regression for single scans or Disease Course Mapping for longitudinal trajectories), encoding this progression information as covariates to the LDM; and iv) Latent Average Stabilization (LAS), a post-hoc technique that averages predictions across multiple inference runs to improve consistency.

The inference process takes a subject's baseline MRI and metadata at age A, predicts progression-related volumetric changes at target age B using the auxiliary model, encodes this information into covariates, and runs the reverse diffusion process conditioned on subject anatomy and progression information to generate a predicted follow-up MRI.

We implemented BrLP using the code and model weights released by the authors. For full architectural and training details, we refer the reader to the original publications(Puglisi et al., 2024; 2025).

A.2.1 EVALUATION PROCEDURE

For each subject with N available visits, we followed the protocol described in the original BrLP publications(Puglisi et al., 2024; 2025): the earliest $\lceil N/2 \rceil$ visits serve as conditioning input, and the remaining visits are held out as prediction targets. Subjects with fewer than two usable visits are excluded. This yielded 163 prediction targets across 150 subjects for OASIS-2 (follow-up intervals up to 6 years) and 4,467 targets across 2,173 subjects for ADNI (follow-up intervals Δt up to 18.1 years).

A.3 BSQ TOKENIZER

A.3.1 DATA SCALING

We experimented with two data scaling strategies: representing the data as the standardized difference from a temporal mean volume ($diff$), or and MinMax scaling, following(Lanzino et al., 2024).

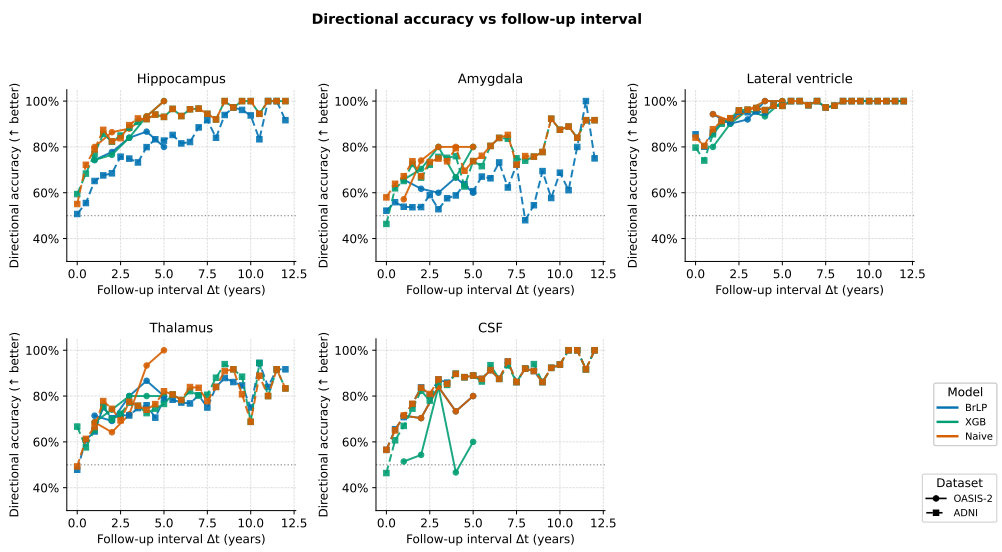

Figure 4: BrLP directional accuracy breakdown as a function of the follow-up interval, compared to the performance of an XGB regressor and a naive guess. Dashed horizontal lines indicate a 50% random baseline. Data was binned in intervals of 0.5 years, bins containing fewer than 10 samples were excluded.

In the former case, we represented the 4D image data I as the sum of a 3D, time-averaged image $\bar{I}$ and a 4D dynamic component:

$$I_{diff} = \frac{I - \bar{I}}{\mathrm{SD(I)}}. \tag{1}$$

Note that here the standard deviation $\mathrm{SD}(I)$ is taken globally, over the entire 4D volume of the brain across all timepoints save for the first 10, to allow the magnetic field and tissue magnetization to stabilize to a steady state. We take the global SD, in contrast with the voxel-wise SD through time, under the assumption that a voxel-wise SD would be more task-dependent compared to the global SD, whereas we aim to train general, task-independent tokenizers.

To further regularize the intensity distribution we take the softsign transform of the signal:

$$I^*_{diff} = \frac{I_{diff}(t)}{1 + |I_{diff}(t)|}. \tag{2}$$

This transformation does not result in any loss of generality as the softsign function is invertible. Conversely, for MinMax normalization:

$$I^*_{MinMax}(t) = \frac{I(t) - \min(I)}{\max(I) - \min(I)}. \tag{3}$$

A.3.2 Architecture

To preserve spatial resolution while enabling discrete tokenization, we designed a convolutional encoder-decoder architecture alternating 3D spatial and temporal convolutions (Figure 5). The encoder applies strided spatial convolutions with kernel size 6 or 12 (to explore different compression factors), followed by alternating spatial and temporal convolution blocks. Each block uses Convolution→SiLU→GroupNorm with residual connections. Channel capacity doubles at each encoder stage before discretization via Binary Spherical Quantization (BSQ)(Zhao et al., 2024). The stride-6 setup featured 64 channels at the first stage, 128-bits for BSQ encoding, and utilized 32 codebooks. These parameters were doubled for the stride-12 tokenizer. The decoder mirrors this structure, halving channels at each stage and reconstructing the original resolution via transpose convolution.

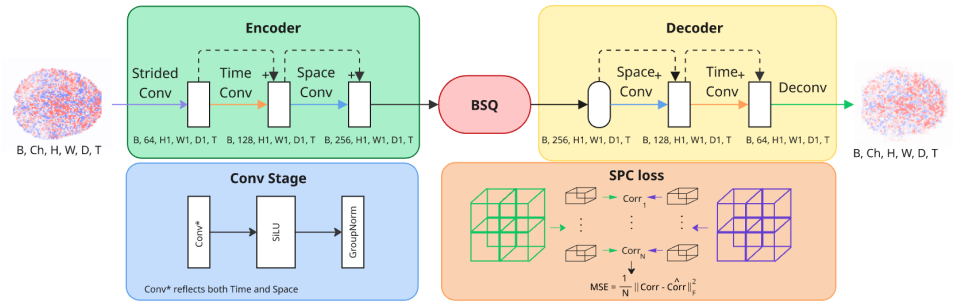

Figure 5: Architecture of the spatiotemporal tokenizer with BSQ bottleneck. Input is a 4D fMRI run (shape (B, Ch, H, W, D, T)) preprocessed using the difference-from-mean strategy. The encoder alternates spatial and temporal convolutions with channel doubling at each stage before BSQ discretization; the decoder mirrors this structure with channel halving. (Bottom Left) Conv Stage detail showing Convolution→SiLU→GroupNorm with residual connections (dashed arrows). (Bottom Right) SPC loss applies random pooling (sizes 3-11) to compute correlation mismatch: $L_{SPC} = \frac{1}{N}\sum(1-\rho_i)^2$. Tensor shapes shown for compression factor 6.

A.3.3 SPC LOSS

We defined the Stochastic Pooling Correlation (SPC) loss to provide a training objective sensitive to temporal dynamics, without requiring pre-defined ROIs or computationally expensive voxel-by-voxel comparisons. For each training batch, we apply average pooling to the original and reconstructed images with random pool sizes from the set $\{3, 4, 5, 7, 11\}$ (Figure 5). By randomly varying the pooling size and grid positioning, we minimize artifacts from fixed pooling patterns. For each pooled region, we extract a timeseries (one value per timepoint), and define ρ_i as the Pearson correlation between pooled timeseries from the original ($l_i(t)$) and reconstructed ($l_i^*(t)$) images. The SPC loss is then:

$$L_{SPC} = \frac{1}{N_{\text{pooled}}} \sum_{i=0}^{N_{\text{pooled}}} (1-\rho_i)^2, \tag{4}$$

which ranges from 0 to 1, penalizing mismatches in temporal correlation. The total tokenizer loss combines quantization, reconstruction, and SPC objectives: $L(I^*, I) = L_{Quant} + rL_{MSE} + sL_{SPC}$ with $r = 50$ and $s = 0.1$.

A.3.4 TRAINING

Tokenizers were trained for 3 epochs using a SGDR strategy(Loshchilov & Hutter, 2016) based on cosine annealing of the learning rate with four warm restarts, using the Adam optimizer with a maximum learning rate of $4 \cdot 10^{-4}$ and a minimum learning rate of $6 \cdot 10^{-5}$. Batch size was 2 with weight decay 10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.99$. The inverse temperature governing the quantization strength for the BSQ module was exponentially annealed from 0.03 to 500 over two epochs. The quantization loss and diversity gamma parameters were annealed linearly over the first epoch from 0 to 1 and 0.67 respectively. Training samples consisted of cubic patches of size 84 in each spatial direction with 9 timepoints; 50 patches were extracted per sample per epoch with randomized ordering across all subjects to prevent test subject leakage.