DreaMR: Diffusion-driven Counterfactual Explanation for Functional MRI

Hasan A. Bedel and Tolga Çukur* *Senior Member*

***Abstract*— Deep learning analyses have offered sensitivity leaps in detection of cognition-related variables from functional MRI (fMRI) measurements of brain responses. Yet, as deep models perform hierarchical nonlinear transformations on fMRI data, interpreting the association between individual brain regions and the detected variables is challenging. Among explanation approaches for deep fMRI classifiers, attribution methods show poor specificity and perturbation methods show limited sensitivity. While counterfactual generation promises to address these limitations, previous counterfactual methods based on variational or adversarial priors can yield suboptimal sample fidelity. Here, we introduce the first diffusion-driven counterfactual method, DreaMR, to enable fMRI interpretation with high fidelity. DreaMR performs diffusion-based resampling of an input fMRI sample to alter the decision of a downstream classifier, and then computes the difference between the original sample and the counterfactual sample for explanation. Unlike conventional diffusion methods, DreaMR leverages a novel fractional multi-phase-distilled diffusion prior to improve inference efficiency without compromising fidelity, and it employs a transformer architecture to account for long-range spatiotemporal context in fMRI scans. Comprehensive experiments on neuroimaging datasets demonstrate the superior fidelity and efficiency of DreaMR in sample generation over state-of-the-art counterfactual methods for fMRI explanation.**

***Index Terms*— counterfactual, explanation, interpretation, generative, diffusion, functional MRI**

I. Introduction

Functional magnetic resonance imaging (fMRI) enables non-invasive cognitive assessments by capturing time-varying blood-oxygen-level-dependent (BOLD) responses across the brain [1]. Spatiotemporally measured BOLD responses can be analyzed to infer associations between individual brain regions and cognition-related variables. A traditional framework in neuroscience performs inference via linear classifiers that are trained to predict variables given responses [2]. The classifier weight for a brain region is then taken to reflect the importance of that region in detecting the respective variable. Although this traditional approach offers ease of interpretation, linear classifiers typically suffer from poor sensitivity [3]. In recent years, deep-learning classifiers have gained prominence as they show substantially higher sensitivity to fine-grained patterns in fMRI data [4]–[12]. Despite this important benefit, hierarchical layers of nonlinear transformation in deep models obscure precise associations between brain responses and cognition-related variables, introducing an interpretation challenge and creating a barrier to methodological trust [13]. As such, there is a dire need for explanation methods that highlight the critically important set of input features (i.e., brain responses) for deep fMRI models to help interpret their decisions.

An emerging framework is counterfactual explanation that aims to identify a minimal, plausible set of changes in the features of an input fMRI sample (i.e., a subject's fMRI scan) that is sufficient to alter the decision of a downstream analysis model [14]. To do this, a generative prior is commonly trained to capture the distribution of original fMRI samples such that new, random samples can be drawn from the learned data distribution [13]. Afterwards, the trained prior is employed to regenerate the values of spatiotemporal responses in an original input sample, in order to produce a counterfactual sample that is proximal to the original sample, albeit that changes the decision of the downstream model [15]. The difference between the two samples is then inspected to interpret associations between brain regions and cognitive state [14]. Counterfactual methods can offer superior feature specificity against attribution methods, which derive gradient or activation heatmaps that can be broadly distributed across input features [16], [17]. They can also produce more sensitive interpretations against perturbation methods, which perform local degradations on input features that can disrupt global coherence [18], [19]. Nevertheless, the performance of counterfactual methods depend critically on the fidelity of samples synthesized by the underlying generative prior.

Previous studies on counterfactual explanation have commonly proposed variational autoencoder (VAE) or generative adversarial network (GAN) priors trained to capture the distribution of fMRI data [14], [15], [20]. These priors synthesize counterfactuals with high efficiency, albeit VAEs often suffer from relatively low sample quality due to loss of detailed features, and GANs suffer from training instabilities that can hamper sample quality or diversity [21]. A promising surrogate for sample generation is diffusion priors that offer high sample fidelity via a many-step sampling process [22], [23]. Few recent imaging studies have considered diffusion-based counterfactual generation to detect anomalous lesions

This study was supported in part by a TUBITAK BIDEB scholarship awarded to H. A. Bedel, by TUBA GEBIP 2015 and BAGEP 2017 fellowships, and by a TUBITAK 1001 Grant 121N029 awarded to T. Çukur (Corresponding author: Tolga Çukur). Funds have been provided by the European Joint Programme Neurodegenerative Disease Research (JPND) 2020 call "Novel imaging and brain stimulation methods and technologies related to Neurodegenerative Diseases" for the Neuripides project 'Neurofeedback for self-stImulation of the brain as therapy for ParkInson Disease'. The Neuripides project has received funding from the following funding organizations under the aegis of JPND: The Netherlands, The Netherlands Organization for Health Research and Development (ZonMw); Germany, Federal Ministry of Education and Research (BMBF); Czech Republic, Ministry of Education, Youth and Sports (MEYS); France, French National Research Agency (ANR); Canada, Canadian Institutes of Health Research (CIHR); Turkey, Scientific and Technological Research Council of Turkey (TUBITAK). H. A. Bedel and T. Çukur are with the Dept. of Electrical and Electronics Engineering, and the National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, Turkey, 06800. (e-mails: {abedel, cukur}@ee.bilkent.edu.tr). T. Çukur is also with the Dept. of Neuroscience, Bilkent University, Ankara, Turkey, 06800.

in anatomical MRI and X-ray scans [24]–[27]. Yet, to our knowledge, diffusion priors have not been explored for counterfactual explanation in multi-variate fMRI analysis. This may be partly related to excessive inference times of conventional diffusion priors such as DDPM or DDIM that require hundreds of steps to generate a single sample [28]. When coupled with the high dimensionality of fMRI data, inefficiency of conventional diffusion priors can be particularly limiting in application domains that benefit from rapid data processing such as real-time fMRI or cohort fMRI studies [29], [30].

Here, we propose a novel diffusion-driven counterfactual explanation method, DreaMR, to interpret downstream fMRI classifiers with improved fidelity and efficiency. The proposed method trains a class-agnostic diffusion prior for fMRI data, and the trained prior generates a counterfactual sample with guidance from a downstream classifier to alter its decision (Fig. 1). To improve inference efficiency without compromising sample quality, DreaMR leverages a novel fractional multi-phase-distilled diffusion (FMD) prior that splits the diffusion process into consecutive fractions and performs multi-phase distillation in each fraction (Fig. 2). Unlike regular diffusion priors implemented with UNet backbone architectures, DreaMR uniquely leverages an efficient transformer architecture of linear complexity in conjunction with the FMD prior to capture long-range spatiotemporal context in fMRI scans. During counterfactual generation, classifier-guidance is injected separately into each diffusion fraction to tailor synthesis of intermediate samples in proximity of the original sample. The difference between the original and the final counterfactual samples reflects the contribution of brain regions to the model decision. We report comprehensive demonstrations to explain deep classifiers for sex detection on resting-state fMRI scans, and for cognitive task detection in task-based fMRI scans. We find that DreaMR achieves superior fidelity against competing explanation methods, and it substantially outperforms conventional diffusion priors in inference efficiency. Code to implement DreaMR is available at https://github.com/icon-lab/DreaMR.

Contributions:

- To our knowledge, we introduce the first diffusion-driven counterfactual explanation method in the literature for multi-variate fMRI analysis.
- DreaMR generates counterfactual samples with a novel fractional multi-phase-distilled diffusion prior to boost inference efficiency without compromising quality.
- Unlike conventional diffusion methods, DreaMR leverages a transformer architecture of linear complexity to capture long-range spatiotemporal context in fMRI scans.

II. Related Work

A. Explanation of fMRI Models

Two prominent frameworks have been proposed in the literature to address the interpretation challenge for deep fMRI classifiers. A first framework is intrinsic interpretation where analyses are conducted using specialized models, such as linearized or graph classifiers, with restricted designs to permit an inherent degree of explanation [11], [31], [32]. Model-specific explanation is then attained by inspecting internal parameters, but use of restricted designs can elicit losses in classification performance [33]–[36]. A more flexible framework is post-hoc explanation where analyses are conducted using an unrestricted model, and the influences of model inputs on the output are observed for interpretation. Among post-hoc techniques, attribution methods derive heatmaps across input features to estimate their salience, taken as gradients [37], [38], activations [39]–[42], or a weighted combination of gradients and activations for the target class [16], [17], [43]. Attribution methods often require architecture-based modifications that limit practicality, and produce over-broad heatmaps that degrade interpretation specificity. To improve local specificity, perturbation methods introduce patch-level degradation on input features through operations such as occlusion [5], [11], [18], [19], [44]–[46]. Yet, perturbation-methods can be computationally heavy and they often produce globally-incoherent results that hamper the sensitivity of interpretations.

Counterfactual generation is an alternative post-hoc technique that resamples original data via a generative prior in order to enforce desired changes in input features [47], [48]. In the case of explaining fMRI classifiers, the aim of counterfactual generation would be to identify a minimally-sufficient set of changes in input features that flip the classifier decision [14]. Although improvements over attribution and perturbation methods have been reported for this application, VAE and GAN priors in previous counterfactual methods can suffer from low sample fidelity that in turn limits the reliability of explanations [13]. Recent machine learning studies advocate diffusion priors as a promising surrogate for reliable sample generation, albeit conventional diffusion priors are known to suffer from an inherent trade-off between sampling quality and efficiency [22], [23]. This has impeded adoption of diffusion priors in counterfactual generation that requires iterated resampling of each high-dimensional fMRI data sample until the respective model decision is flipped. To address these open issues, here we introduce DreaMR, the first diffusion-driven explanation method for multivariate fMRI analysis to our knowledge.

B. Counterfactual Generation

Recent computer vision studies have considered diffusion priors to generate counterfactual natural images from desired object classes. [49]–[51] propose conventional diffusion priors based on a common UNet architecture. Interleaved sampling is used to trade-off sample quality in return for efficiency [50], [51]. Unlike conventional diffusion priors that typically require hundreds of diffusion steps even with interleaved sampling, DreaMR improves practicality via its novel FMD prior that achieves high sampling efficiency and quality via multi-phase distillation on consecutive fractions of the diffusion process, where each fraction uses a dedicated denoising network. To our knowledge, FMD is the first diffusion prior to perform fraction-specific distillation in the literature. Furthermore, DreaMR uniquely implements denoising networks via a transformer architecture to improve sensitivity to long-range context in fMRI scans that last several minutes [52].

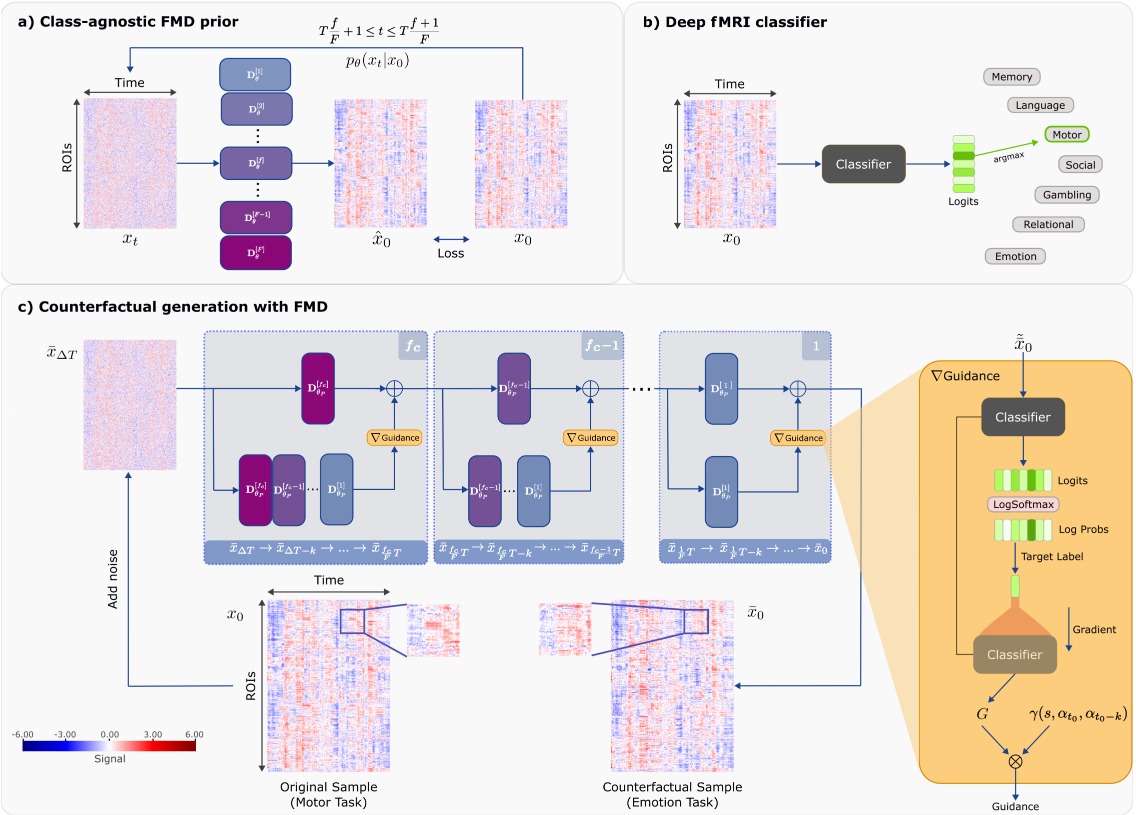
a) Class-agnostic FMD prior
$T\frac{f}{F}+1 \leq t \leq T\frac{f+1}{F}$
$p_\theta(x_t|x_0)$
Time
ROIs
x_t
$\mathbf{D}_\theta^{[1]}$
$\mathbf{D}_\theta^{[2]}$
$\mathbf{D}_\theta^{[f]}$
$\mathbf{D}_\theta^{[F-1]}$
$\mathbf{D}_\theta^{[F]}$
$\hat{x}_0$
Loss
x_0
b) Deep fMRI classifier
Time
ROIs
x_0
Classifier
Logits
argmax
Memory
Language
Motor
Social
Gambling
Relational
Emotion
c) Counterfactual generation with FMD
$\bar{x}_{\Delta T}$
f_c
f_c-1
1
∇Guidance
$\bar{x}_{\Delta T} \rightarrow \bar{x}_{\Delta T-k} \rightarrow \dots \rightarrow \bar{x}_{\frac{f_c}{F}T}$
$\bar{x}_{\frac{f_c}{F}T} \rightarrow \bar{x}_{\frac{f_c}{F}T-k} \rightarrow \dots \rightarrow \bar{x}_{\frac{f_c-1}{F}T}$
$\bar{x}_{\frac{1}{F}T} \rightarrow \bar{x}_{\frac{1}{F}T-k} \rightarrow \dots \rightarrow \bar{x}_0$
Add noise
Time
ROIs
x_0
$\bar{x}_0$
-6.00 -3.00 0.00 3.00 6.00
Signal
Original Sample (Motor Task)
Counterfactual Sample (Emotion Task)
$\tilde{\bar{x}}_0$
Classifier
Logits
LogSoftmax
Log Probs
Target Label
Gradient
Classifier
G
$\gamma(s, \alpha_{t_0}, \alpha_{t_0-k})$
Guidance

Fig. 1: DreaMR is a counterfactual explanation method for deep fMRI classifiers. **a)** To capture the distribution of fMRI data, DreaMR trains a class-agnostic FMD prior that splits the diffusion process into F uniform fractions with dedicated networks $\mathbf{D}_\theta^{[f]}$, as described in Eq. 6. Following training, the FMD prior is subjected to multi-phase distillation to attain $\mathbf{D}_{\theta_P}^{[f]}$ that allow fast sampling, as described in Eq. 7. **b)** DreaMR is devised to explain the decisions of a deep classifier that predicts cognition-related variables from a subject's fMRI scan. **c)** Given an input fMRI sample $x_0 \sim p(x)$ mapped onto cognitive state y_0 by the classifier, DreaMR first samples a noise-added version $\bar{x}_{\Delta T}$ via forward diffusion and then generates $\bar{x}_0 \sim p(x)$ with minimal alterations from x_0 via reverse diffusion. Generation is guided with the conditional score of the classifier to flip the decision to $\bar{y}_0 \neq y_0$. Starting at fraction $f_c = \lceil \Delta T(F/T) \rceil$, classifier guidance computed from a denoised estimate of the counterfactual sample ($\tilde{\bar{x}}$) is injected at each diffusion step (orange boxes), as described in Alg. 1.

Few recent imaging studies have also considered diffusion-based counterfactual generation to map lesions in anatomical scans by synthesizing pseudo-healthy medical images [24]–[26]. Commonly, these studies propose conventional diffusion priors based on UNet. Instead, DreaMR leverages the novel FMD prior for higher efficiency and a transformer architecture to capture long-range context. Note that [24], [25] train class-conditional priors on normals and patients, and [26] trains a class-specific prior on normals. When adopted for counterfactual explanation, such class-informed priors must use matching class definitions to the classifier, so they require retraining for each classification task. In contrast, DreaMR employs a class-agnostic diffusion prior that can be utilized to explain models for different tasks without retraining.

III. Theory

A. Conventional Diffusion Priors

Diffusion priors use a gradual process to transform between a data sample x_0 and a random noise sample x_T in T steps. In the forward direction, Gaussian noise is added to obtain a noisier sample x_t at step t, with forward transition probability:

$$q\left(x_t|x_{t-1}\right) = \mathcal{N}\left(x_t; (\alpha_t/\alpha_{t-1})x_{t-1}, \sigma^2_{t|t-1}\mathrm{I}\right), \quad (1)$$

where $\mathcal{N}$ is a Gaussian distribution, I is the identity covariance matrix, and α, σ^2 are scaling and noise variance parameters where $\alpha_t^2 + \sigma_t^2 = 1$ [22]. In the reverse direction, a network parametrization $\mathbf{D}_\theta$ is used to restore original data from noisy samples, i.e., $\hat{x}_0 = \mathbf{D}_\theta(x_t, t)$. The prior can be trained by minimizing a variational bound on likelihood [23]:

$$\mathbb{E}_{t\sim U[1,T], x_0 \sim p(x), x_t \sim q(x_t|x_0)} \left[\omega\left(\lambda_t\right) \|\mathbf{D}_\theta(x_t,t) - x_0\|_2^2\right], \quad (2)$$

where $\mathbb{E}$ is expectation, U is a uniform distribution, $q(x_t|x_0) = \mathcal{N}\left(x_t; \alpha_t x_0, \sigma_t^2 \mathrm{I}\right)$, $\omega\left(\lambda_t\right)$ is a weighting function with $\lambda_t = log(\alpha_t^2/\sigma_t^2)$ denoting signal-to-noise ratio.

Once trained, the diffusion prior can generate a synthetic data sample by progressively denoising a random noise sample across T steps. The reverse transition probability for the denoising steps can be expressed as [22]:

$$q\left(x_{t-1}|x_t, \hat{x}_0\right) = \mathcal{N}\left(\alpha_{t-1}\hat{x}_0 + \sqrt{1-\alpha_{t-1}^2-\beta_t^2}\,.\frac{x_t - \alpha_t \hat{x}_0}{\sigma_t},\ \beta_t^2 \mathrm{I}\right) (3)$$

where $\beta_t = \frac{\sigma_{t-1}}{\sigma_t\ \alpha_{t-1}}\sqrt{(\alpha_{t-1}^2 - \alpha_t^2)}$ controls the stochasticity of generated samples. The original diffusion prior requires $T\approx 1000$ forward passes through the network for generation. To lower sampling time, a common solution is interleaved sampling with step size k, $x_{t-k} \sim q(x_{t-k}|x_t, \hat{x}_0)$, while $\beta_t = 0$ for deterministic generation (i.e., DDIM) [28]. Yet, interleaved sampling can still require few hundred steps for generation and it typically suffers from reduced sample quality [50].

B. DreaMR

DreaMR is a novel explanation method for deep fMRI models based on counterfactual generation. Assume that a downstream classifier $c(x){=}y$ maps an input fMRI sample $x_0{\sim}p(x)$ onto a class label $y_0 \in Y$ for cognition-related variable, according to posterior probability $p_c(y|x)$. Here, we take that $x_0 \in \mathbb{R}^{R\times W}$ denotes the BOLD responses recorded in the given fMRI scan, where R is the number of brain regions and W is the number of time frames. Counterfactual generation aims to obtain plausible samples $\bar{x}_0{\sim}p(x)$ with minimal alterations from x_0, such that the classifier decision is flipped $c(\bar{x}_0) = \bar{y}_0$ where $\bar{y}_0 \neq y_0, \bar{y}_0 \in Y$. Afterwards, the differences between the original and counterfactual samples (x_0–$\bar{x}_0$) can be inspected to infer the input features that are critical in distinguishing between labels y_0 and $\bar{y}_0$.

Counterfactual generation inherently requires alteration of response values within the input fMRI sample. To do this based on a trained diffusion prior, a noisy fMRI sample $\bar{x}_{\Delta T}$ is first obtained by adding a moderate level of white Gaussian noise onto the original sample x_0 [49]:

$$q\left(\bar{x}_{\Delta T}|x_0\right) = \mathcal{N}\left(\bar{x}_{\Delta T}; (\alpha_{\Delta T}/\alpha_0)x_0, \sigma^2_{\Delta T|0}\mathrm{I}\right), \quad (4)$$

where the time step $\Delta T < T$ is a hyperparameter. Starting reverse diffusion at ΔT, response values in the noisy fMRI sample can then be altered according to the reverse transition probabilities in Eq. 3, so that the resultant counterfactual sample remains proximal to x_0. Yet, to ensure that the counterfactual sample is able to flip the classifier decision, classifier guidance is also injected [50]:

$$q(\bar{x}_{t-1}|\bar{x}_t, \hat{\bar{x}}_0) = \mathcal{N}\left(\hat{\bar{x}}_0 + s\,\beta_t^2\,\nabla_{\bar{x}_t} log\, p_c(\bar{y}_0|\bar{x}_t), \beta_t^2\right), \quad (5)$$

where $\nabla_{\bar{x}_t}$ is gradient with respect to $\bar{x}_t$, s is a scaling parameter that controls the relative weight of classifier guidance.

Previous studies on counterfactual generation have adopted interleaved sampling with conventional diffusion priors, which still takes few hundred sampling steps [28]. Furthermore, counterfactual generation requires knowledge of $\nabla_{\bar{x}_t} log\, p_c(\bar{y}|\bar{x}_t)$ that is unknown a priori as the classifier is trained on original samples without added noise. To improve inference efficiency, DreaMR leverages a novel FMD prior for counterfactual generation in few steps without compromising sample quality (Fig. 1). Meanwhile, to avoid the need for classifier retraining on noisy samples, DreaMR computes classifier gradients given denoised sample estimates as a surrogate for gradients on noisy samples. Working principles of the FMD prior, classifier guidance and the counterfactual generation algorithm are detailed in the rest of this section.

B.1 Fractional multi-phase-distilled diffusion prior: During counterfactual generation with diffusion priors, providing guidance via classifier gradients at each diffusion step can cause significant computational burden for inference [24]. A mainstream approach to improve inference efficiency is distillation for post-hoc reduction of the number of diffusion steps [49], [53]. However, conventional distillation procedures on common diffusion priors can result in undesirable losses in sample quality [54], [55]. Here, we argue that two main contributors to these losses are the use of a single denoising network for the entire diffusion process and the use of a single-phase distillation. Thus, to enable efficient inference without compromising sample quality, we propose a novel FMD prior that splits the diffusion process into F uniform fractions with dedicated denoising networks, and performs multi-phase distillation separately in each fraction (Fig. 2).

Characteristics of the denoising task can show notable variations within a diffusion process due to varying noise levels and feature details across diffusion steps [56]. In turn, poor adaptation to these characteristics can induce significant performance losses in distilled diffusion priors. Several recent studies have considered to employ multiple denoising networks on separate time fractions to improve adaptation in undistilled diffusion priors [56], but these studies did not examine the influence of fractional diffusion on distilled priors. While other studies have considered progressive distillation over multiple stages to alleviate losses in distilled priors [54], they employed a common denoising network across the entire diffusion process that can compromise sample quality. Unlike these recent efforts, here we introduce FMD as the first diffusion prior that synergistically combines fractional diffusion with fraction-specific multi-phase distillation to simultaneously maintain high quality and efficiency in sample generation.

For fractional diffusion, FMD employs a dedicated denoising network $\hat{x}_0 = \mathbf{D}_\theta^{[f]}(x_t, t)$ in each fraction (Fig. 2). The fth fraction covers T/F consecutive steps from $t_s(f) = T(f)/F$ to $t_e(f) = T(f-1)/F+1$. The resultant FMD prior is trained via the following objective:

$$\sum_{f=1}^{F} \mathbb{E}_{t\sim U[t_s(f), t_e(f)]}\left[\|\mathbf{D}_\theta^{[f]}(x_t, t) - x_0\|_2^2\right], \quad (6)$$

where ω and expectation over x_0, x_t are omitted for brevity.

For multi-phase distillation, FMD performs gradual knowledge transfer from the original teacher network $\mathbf{D}_{\theta_0}^{[f]}(x_t, t)$ onto a student network $\mathbf{D}_{\theta_P}^{[f]}$ over P phases. In the pth phase, $\mathbf{D}_{\theta_{p-1}}^{[f]}(x_t, t)$ is the teacher, $\mathbf{D}_{\theta_p}^{[f]}(x_t, t)$ is the student, and the diffusion step size is increased by a factor of 2 as follows:

$$\mathbb{E}_{t\sim U(\{t_s(f):T/(k_d F):t_e(f)\})}\left[\|\mathbf{D}_{\theta_p}^{[f]}(x_t, t) - \tilde{x}_0\|_2^2\right], \quad (7)$$

where $k_d = 2^p$ is the diffusion step size of the student. Adopting interleaved sampling, the reference sample $\tilde{x}_0$ is derived via sampling based on the teacher $\mathbf{D}_{\theta_{p-1}}^{[f]}(x_t, t)$:

$$\tilde{x}_0 = \frac{\tilde{x}_{t-2k_o} - (\sigma_{t-2k_o}/\sigma_t)x_t}{\alpha_{t-2k_o} - (\sigma_{t-2k_o}/\sigma_t)\alpha_t}, \quad (8)$$

where $k_o = 2^{(p-1)}$ is the diffusion step size of the teacher. At the end of the distillation procedure, the number of steps for a given fraction is reduced from T/F to $T/(2^P F)$. Although this multi-phase distillation involves additional computations over a single-phase distillation during the training stage, progressively lowering the number of diffusion steps helps mitigate losses in sample quality in the distilled diffusion prior.

B.2 Classifier guidance: A counterfactual sample that flips the classifier decision can be drawn if the joint score function $\nabla_{x_t} log\, p(x_t, y)$ for noisy samples and predicted class labels is known. Since $p(x_t, y) = p(x_t)p_c(y|x_t)$ based on Bayes' rule, the joint score is given as:

$$\nabla_{x_t} log\, p(x_t, y) = \nabla_{x_t} log\, p(x_t) + \nabla_{x_t} log\, p_c(y|x_t). \quad (9)$$

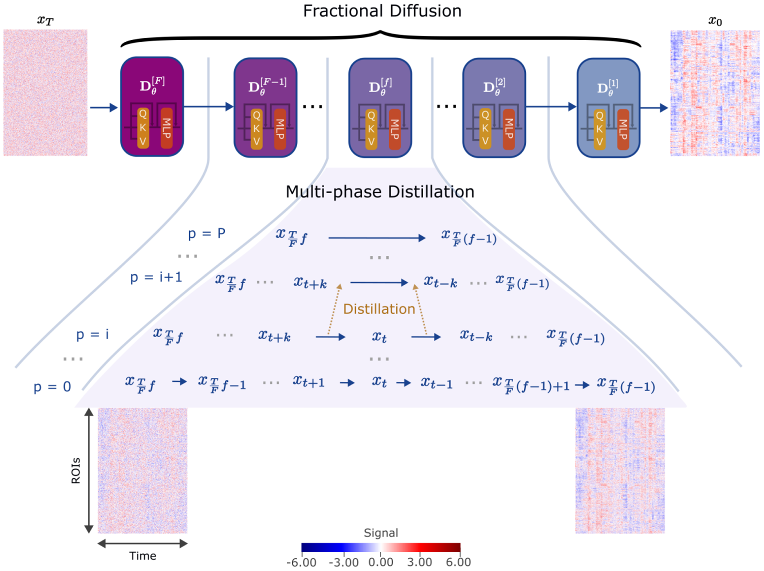

Fig. 2: DreaMR leverages a novel FMD prior for efficient sample generation without compromising sample quality, and to enable effective intermittent control over the generation process. The FMD prior splits the overall diffusion process into F fractions, where the fth fraction covers T/F consecutive steps from $t_s(f) = T(f)/F$ to $t_e(f) = T(f-1)/F + 1$. Since the characteristics of the denoising task can vary notably across fractions, a dedicated denoising network $\mathbf{D}_{\theta_P}^{[f]}$ is employed in each fraction to improve sample fidelity. To enhance sensitivity to long-range temporal context in fMRI scans, the network is built on an efficient transformer architecture that uses fused window attention mechanisms [36]. To improve sampling efficiency, multi-phase distillation is performed over P phases in each fraction (Eq. 7). In the pth distillation phase, the diffusion step size is doubled to shorten sampling time by a factor of 2. At the end of multi-phase distillation, the number of sampling steps that must be executed for a given fraction is reduced from T/F to $T/(2^P F)$. Note that the novel combination of diffusion fractions and multi-phase distillation in FMD serves to mitigate losses in sample quality typically encountered with distillation of conventional diffusion priors.

The first term denotes the marginal score function for noisy samples, which can be derived as follows [22]:

$$p(x_t) = \frac{1}{\sigma_t\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x_t - \alpha_t\hat{x}_0}{\sigma_t}\right)^2\right), \quad (10)$$

$$\nabla_{x_t} log\, p(x_t) = -\frac{x_t - \alpha_t\hat{x}_0}{\sigma_t^2}. \quad (11)$$

Meanwhile, the second term denotes the conditional score function for predicted labels given noisy samples. Note that the originally trained classifier does not capture $p_c(y|x_t)$, but instead $p_c(y|x_0)$. The classifier could be retrained on a set of noisy samples generated by the diffusion prior. However, this brings in additional computational burden, and learning of $p_c(y|x_t)$ on samples with heavy noise might be difficult.

To avoid limitations related to classifier retraining, here we derive a surrogate based on $p_c(y|x_0)$ to compute the conditional score function. Following the chain rule:

$$\nabla_{x_t} log\, p_c(y|x_t) = \nabla_{\tilde{x}_0} log\, p_c(y|x_t) \cdot \nabla_{x_t}\tilde{x}_0, \quad (12)$$

where $\tilde{x}_0$ is obtained via sampling across the diffusion process. Based on forward diffusion, x_t and x_0 are related as:

$$x_t = \alpha_t x_0 + \sigma_t\epsilon, \text{ s.t. } \epsilon \sim \mathcal{N}(0, \mathrm{I}). \quad (13)$$

Assuming that $\tilde{x}_0 \simeq x_0$, $\nabla_{x_t}\tilde{x}_0 = 1/\alpha_t$. In conventional diffusion priors with stochastic sampling (i.e., $\beta_t \neq 0$), x_t and $\tilde{x}_0$ can be related through a one-to-many mapping. Yet, since DreaMR adopts deterministic sampling with $\beta_t = 0$, a unique $\tilde{x}_0$ is obtained given x_t. As such, $p_c(y|x_t) = p_c(y|\tilde{x}_0)$, and the conditional score can be expressed as:

$$\nabla_{x_t} log\, p_c(y|x_t) = \frac{1}{\alpha_t}\nabla_{\tilde{x}_0} log\, p_c(y|\tilde{x}_0). \quad (14)$$

Since $p_c(y|\tilde{x}_0) \simeq p_c(y|x_0)$ for a reasonably well-trained diffusion prior, the posterior distribution of the originally trained classifier can be used to provide guidance without retraining.

B.3 Counterfactual generation: The counterfactual generation algorithm for DreaMR is outlined in Alg. 1. Given an original fMRI sample x_0, a noisy sample $\bar{x}_{\Delta T}$ is first obtained by adding Gaussian noise via forward diffusion as described in Eq. 13. DreaMR then performs reverse diffusion sampling via the trained FMD prior to alter the response values in $\bar{x}_{\Delta T}$, without diverging significantly from x_0. Note that time step ΔT corresponds to fraction $f_c = \lceil \Delta T(F/T) \rceil$, so reverse diffusion is initiated at the f_cth fraction. To ensure that the eventual counterfactual sample is able to flip the classifier decision to $\bar{y}$, gradient of the posterior distribution of the classifier is also employed. For the fth fraction, the interleaved sampling equation based on the FMD prior injected with classifier guidance can be described as:

$$\begin{aligned} \bar{x}_{t-k} \quad &= \quad \alpha_{t-k}\hat{\bar{x}}_0 + \sigma_{t-k}\frac{\bar{x}_t - \alpha_t\hat{\bar{x}}_0}{\sigma_t} + \\ & s\frac{\sigma_t^2}{\alpha_t^2}(\alpha_{t-k} - \alpha_t\frac{\sigma_{t-k}}{\sigma_t})\nabla_{\tilde{\bar{x}}_0} log\, p_c(\bar{y}|\tilde{\bar{x}}_0), \quad (15) \end{aligned}$$

where k is step size, $\hat{\bar{x}}_0 = \mathbf{D}_{\theta_P}^{[f]}(\bar{x}_t, t)$, and s is a scaling constant to control the strength of guidance. In Eq. 15, the conditional score from the classifier is computed using $\tilde{\bar{x}}_0$, which is Langevin sampled across the diffusion process, instead of $\hat{\bar{x}}_0$, as this was observed to improve the quality of guidance. The counterfactual sample $\bar{x}_0$ is obtained after taking $\Delta T/k$ reverse diffusion steps.

IV. Methods

A. Experimental Procedures

Demonstrations were performed on fMRI scans from HCP-Rest, HCP-Task [57] and ID1000 datasets [58]. After exclusion of incomplete scans (<1200s), HCP-Rest contained 1093 resting-state fMRI samples (594 female, 499 male). HCP-Task contained 7450 task-based fMRI samples (594 female, 501 male) where each subject performed 7 different tasks (i.e., emotion, relational, gambling, language, social, motor, working memory). ID1000 contained 881 movie-watching fMRI samples (458 female, 423 male).

As downloaded from the public datasets, several preprocessing steps had already been performed on fMRI scans. For HCP-Rest and HCP-Task, motion correction, distortion correction, registration onto the MNI template, bias field correction, and brain extraction had been performed [59]. For ID1000, motion correction, distortion correction, registration onto the ICBM template, brain extraction, and CompCorr denoising

Algorithm 1: Counterfactual generation with DreaMR

```
Input:
$x_0 \sim p(x)$: Original fMRI sample
$p_c(y|x)$: Posterior probability of the classifier
$y_0 \in Y$: Classifier-predicted label for $x_0$
$\bar{y}_0$: Target label for counterfactual generation
$\Delta T$: Initial diffusion step for counterfactual generation
$\{\mathbf{D}_{\theta_P}^{[1]}, ..., \mathbf{D}_{\theta_P}^{[F]}\}$: Distilled networks across fractions
$k = 2^P$: Step size after P distillation phases
Output:
$\bar{x}_0$: Counterfactual sample

$\bar{x}_{\Delta T} \leftarrow \alpha_{\Delta T} x_0 + \sigma_{\Delta T}\epsilon$;  ▷ generate noisy sample
for $f_o$ in range($f_c$, 0, -1) do
    ▷ sample across fractions to find $\bar{\bar{x}}_0$
    for $t_i$ in range($t_s(f_o)$, $-1$, $-k$) do
        $f_i \leftarrow ceil(t_i F/T)$;
        $\hat{\bar{x}}_0 \leftarrow \mathbf{D}_{\theta_P}^{[f_i]}(\bar{x}_{t_i}, t)$;
        $\bar{x}_{t_i-k} \leftarrow \alpha_{t_i-k}\hat{\bar{x}}_0 + \sigma_{t_i-k}\frac{\bar{x}_{t_i} - \alpha_{t_i}\hat{\bar{x}}_0}{\sigma_{t_i}}$;
    end
    $\bar{\bar{x}}_0 \leftarrow \bar{x}_0$; $G \leftarrow \nabla_{\bar{\bar{x}}_0} log\, p_c(\bar{y} \,|\, \bar{\bar{x}}_0)$; ▷ gradient
    ▷ sample within fraction to find $\bar{x}_{t_e(f)}$
    for $t_o$ in range($t_s(f_o)$, $t_e(f_o) - 1$, $-k$) do
        $\hat{\bar{x}}_0 \leftarrow \mathbf{D}_{\theta_P}^{[f_o]}(\bar{x}_{t_o}, t)$;
        $\gamma \leftarrow s\frac{\sigma_{t_o}^2}{\alpha_{t_o}^2}(\alpha_{t_o-k} - \alpha_{t_o}\frac{\sigma_{t_o-k}}{\sigma_{t_o}})$; ▷ scale
        $\bar{x}_{t_o-k} \leftarrow \alpha_{t_o-k}\hat{\bar{x}}_0 + \sigma_{t_o-k}\frac{\bar{x}_{t_o} - \alpha_{t_o}\hat{\bar{x}}_0}{\sigma_{t_o}} + \gamma G$;
    end
end
```

had been performed [58]. In addition to these steps, we created nuisance variables for motion and physiological noise due to respiratory and cardiac traces across each fMRI run [60]. To control for confounds, these nuisance variables were regressed out from voxel-wise fMRI time series, and the resultant time series were z-scored across time for normalization.

Regions of interest (ROI) in the brain were defined based on the Schaefer atlas (R=400 regions) [61], the MMP atlas (R=360 regions) [62], or an aggregate atlas that pooled MMP with 20 cerebellar and 24 subcortical ROIs from the Talairach atlas (R=404) [63]. Responses within each ROI were obtained by averaging fMRI signals across member voxels according to the ROI definition. As such, each fMRI scan was represented via a data matrix of $R \times W$, with the number of time frames W=1200 for HCP-Rest, W=176-405 for HCP-Task depending on cognitive task, and W=290 for ID1000. To cope with varying scan durations, a sliding window approach was adopted across the time dimension of fMRI scans [36]. Counterfactual methods were devised to process individual strided windows across the scans, and window-specific outputs were merged by averaging overlapping time frames between consecutive windows. The window size was 600 for HCP-Rest, 128 for HCP-Task and ID1000, and the stride was taken as half the window size [36].

Experiments were conducted on single NVIDIA RTX 3090 GPUs using the PyTorch framework. Modeling was performed via a five-fold cross-validation procedure. In each fold, data were randomly split into training (80%), validation (10%) and test sets (10%) without any subject overlap among the three sets. Data splits were devised to ensure that there was no subject overlap among the validation sets, or among the test sets for separate folds. For fair comparison, identical data splits were used for all competing methods. Model performances were evaluated on the test set for each cross-validation fold, and reported as mean±std across five folds.

A transformer-based downstream fMRI classifier was trained using cross-entropy loss [36]. For explanation methods, priors were trained using their originally proposed losses, and hyperparameters including number of epochs, learning rate and batch size were selected to minimize validation loss [64]. For each method, a common set of hyperparameters yielding near-optimal performance was used across datasets.

B. Competing Methods

DreaMR was compared against state-of-the-art counterfactual methods based on VAE, GAN, and diffusion priors.

B.1 DreaMR: DreaMR was implemented based on the transformer architecture in [36], adapted for diffusion modeling by incorporating time encoding via adaptive normalization layers [65]. Inspired by efficient transformer methods based on temporal windowing [66], [67], the transformer processed time series across a hierarchically growing set of time windows to maintain linear complexity. The FMD prior used T=1024 steps, F=4 fractions, P=7 distillation phases, k=128 final step size. Cross-validated hyperparameters were E=100 epochs, η=2x10^{-4} learning rate, B=8 batch size, s=(8,16,32) for HCP-Task, ID1000, HCP-Rest respectively.

B.2 DiME: A diffusion-based method was trained to generate fMRI samples; counterfactuals were generated by interleaved sampling guided by downstream classification loss and perceptual loss between original and resampled fMRI scans [51]. E=200, η=5x10^{-5}, B=8 were cross-validated.

B.3 DiffSCM: A diffusion-based method was trained to generate fMRI samples; counterfactuals were generated by interleaved sampling guided by downstream classification loss [50]. E=200, η=10^{-4}, B=8 were cross-validated.

B.4 CAG: A GAN-based method was trained to translate between classes given cross-entropy loss from the downstream classifier; counterfactuals were generated via translation [14]. E=200, η=5x10^{-5}, B=32 were cross-validated.

B.5 CheXplain: A GAN-based method was trained to generate samples given predictions by the downstream classifier; counterfactuals were generated by modifying the style latents [68]. E=200, η=2x10^{-4}, B=16 were cross-validated.

B.6 DiVE: A VAE-based method was trained via a variational objective to generate fMRI samples; counterfactuals were generated by modifying encoded latents of the autoencoder [15]. E=500, η=4x10^{-4}, B=16 were cross-validated.

B.7 LatentShift: An autoencoder-based method was considered that modified latent representations of an input fMRI sample to emphasize features that contribute to the classifier decision [20]. E=100, η=4x10^{-4}, B=16 were cross-validated.

B.8 D'artagnan: A GAN-based method was trained to translate between classes given cross-entropy loss from the downstream classifier and distance between the original and translated fMRI scans; counterfactuals were generated via translation [47]. E=100, η=1x10^{-4}, B=8 were cross-validated.

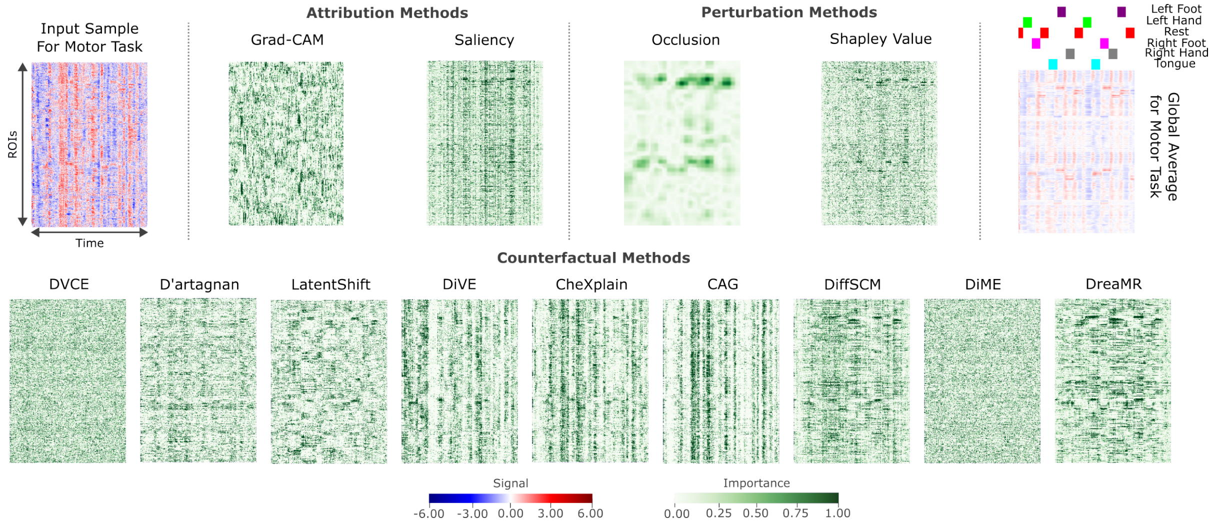

Fig. 3: Spatiotemporal explanation maps from competing methods for a representative fMRI scan in HCP-Task based on the Schaefer atlas. The original input fMRI sample for the motor task is shown on the left. The global average of fMRI samples across subjects for the motor task is shown on the right, where colored boxes are used to depict the time slots within fMRI scans during which specific motor tasks were performed (left foot: purple, left hand: green, . . . , tongue: cyan). For attribution methods, explanation maps were taken as the gradient of the classifier loss function with respect to input features. For perturbation methods, explanation maps were obtained by masking out local patches in fMRI samples. For counterfactual methods, counterfactuals were generated separately to flip the class label from the motor onto each of six remaining cognitive tasks, and explanation maps were taken as the average difference between the original and counterfactual samples.

TABLE I: Fidelity of spatiotemporal features of counterfactual samples was measured via proximity (Prox.), sparsity (Spar.), and FID metrics. Results listed for HCP-Rest, HCP-Task and ID1000 datasets based on the Schaefer atlas, as mean±std across five cross-validation folds. Boldface marks the top-performing method in each dataset.

	HCP-Rest			HCP-Task			ID1000		
	Prox. ↓	Spar. ↓	FID ↓	Prox. ↓	Spar. ↓	FID ↓	Prox. ↓	Spar. ↓	FID ↓
DreaMR	**42.0±0.5**	**12.0±0.2**	21.3±0.4	**49.7±0.5**	15.2±0.2	**5.0±0.4**	**50.1±0.4**	15.6±0.2	**11.5±0.3**
DiME	64.8±1.3	20.4±0.5	25.7±1.6	65.5±0.6	20.8±0.2	12.8±0.9	89.2±0.7	28.7±0.2	16.8±0.3
DiffSCM	64.2±4.9	14.4±0.9	40.9±2.3	66.6±5.7	**14.5±1.3**	17.0±1.1	74.8±11.3	**14.4±1.0**	25.4±6.8
CAG	152.6±46.5	39.1±6.0	**21.2±5.4**	185.0±41.2	43.9±5.7	9.8±1.8	189.0±5.2	45.7±0.7	13.5±1.5
CheXplain	143.2±5.7	39.5±0.9	208.3±54.8	162.6±7.7	42.5±1.1	72.3±11.5	161.5±10.0	42.5±1.3	91.2±21.7
DiVE	101.6±1.6	30.9±0.4	139.3±4.5	107.7±0.8	32.6±0.2	124.6±3.7	159.1±0.7	42.1±0.1	262.8±7.9
LatentShift	49.5±1.0	14.6±0.3	69.1±1.0	87.1±18.8	25.2±3.7	170.7±3.0	118.8±25.2	33.9±5.1	172.8±3.7
D'artagnan	123.5±30.8	34.5±4.9	122.6±24.8	80.3±6.9	25.0±1.9	61.7±13.0	126.7±26.0	36.3±4.3	77.6±26.9
DVCE	67.3±5.5	20.6±1.2	26.2±1.1	58.0±0.5	18.2±0.2	20.6±1.7	68.4±0.5	22.3±0.2	20.5±0.9

TABLE II: Fidelity of spatiotemporal features of counterfactual samples for HCP-Rest, HCP-Task and ID1000 based on the MMP atlas.

	HCP-Rest			HCP-Task			ID1000		
	Prox. ↓	Spar. ↓	FID ↓	Prox. ↓	Spar. ↓	FID ↓	Prox. ↓	Spar. ↓	FID ↓
DreaMR	**38.4±0.5**	**10.5±0.5**	20.0±0.6	**49.4±0.6**	15.0±0.2	**4.9±0.4**	**51.4±0.6**	16.1±0.2	12.9±0.5
DiME	54.6±1.6	16.7±0.6	29.5±1.9	62.6±1.4	19.6±0.5	9.7±1.9	88.3±0.5	28.4±0.2	16.8±0.4
DiffSCM	50.2±2.7	10.8±0.6	39.5±2.9	52.5±3.0	**11.0±0.5**	13.9±1.4	74.7±8.1	**13.7±1.3**	29.3±3.6
CAG	181.9±49.0	42.6±6.3	**19.0±1.3**	184.4±53.6	43.3±7.4	9.6±0.5	195.6±15.7	45.9±1.9	**11.0±0.3**
CheXplain	158.5±27.0	41.3±3.6	242.4±51.9	174.3±12.5	43.8±1.7	67.0±15.3	181.1±16.0	45.1±1.9	75.0±22.8
DiVE	99.0±4.9	30.2±1.1	129.8±8.9	97.8±0.6	30.3±0.2	112.8±2.5	150.2±0.4	41.1±0.1	139.1±8.4
LatentShift	40.9±0.6	11.3±0.2	41.8±1.1	91.7±21.5	26.5±3.4	98.7±23.4	100.2±10.0	30.3±2.4	108.6±5.0
D'artagnan	69.8±2.6	21.9±0.9	65.2±16.8	56.9±18.0	17.1±5.8	27.0±5.7	93.8±31.4	28.4±6.7	39.1±18.2
DVCE	46.6±1.5	13.7±0.6	30.7±2.2	54.8±0.6	17.0±0.2	16.3±2.4	67.0±0.4	21.9±0.2	19.9±0.8

B.9 DVCE: A diffusion-based method was trained to generate fMRI samples; counterfactuals were obtained by interleaved sampling guided by downstream classification loss [49]. E=100, η=5x10^{-5}, B=8 were cross-validated.

C. Performance Evaluation

We quantified proximity, sparsity and Fréchet Inception Distance (FID) metrics to assess performance of competing counterfactual methods [49]. Proximity was taken as the normalized ℓ_2 distance between original and counterfactual samples; sparsity was taken as the proportion of features whose absolute difference between original and counterfactual samples exceeded the standard deviation across features in the original sample. Low proximity and sparsity indicate that feature changes between samples are minimal, so the resultant interpretations are specific. Meanwhile, low FID scores suggest that counterfactual samples are drawn from a similar distribution to that of original samples, so the interpretations are plausible. Significance of differences between methods were assessed via non-parametric Wilcoxon signed-rank tests. FID produced an aggregate measure across the test set so it was not included in significance assessments.

TABLE III: Fidelity of functional connectivity (FC) features of counterfactual samples for HCP-Rest, HCP-Task and ID1000 based on the Schaefer atlas.

	HCP-Rest			HCP-Task			ID1000		
	Prox. ↓	Spar. ↓	FID ↓	Prox. ↓	Spar. ↓	FID ↓	Prox. ↓	Spar. ↓	FID ↓
DreaMR	**0.6±0.0**	**3.5±0.5**	**41.4±2.6**	**0.9±0.0**	**5.5±0.4**	**11.3±1.0**	**0.8±0.0**	**2.8±0.1**	**33.0±1.3**
DiME	0.9±0.1	7.8±1.1	48.1±3.2	1.8±0.1	17.5±1.2	26.2±1.4	1.6±0.0	10.2±0.1	54.3±2.3
DiffSCM	2.9±0.4	26.8±2.3	120.6±4.9	4.7±0.3	30.4±2.0	64.2±4.6	2.0±0.3	14.9±1.4	76.7±4.1
CAG	3.5±0.3	35.7±2.0	53.5±8.7	3.9±0.2	33.4±1.6	21.1±7.5	4.0±0.1	28.8±0.5	38.4±6.7
CheXplain	13.9±2.2	65.7±3.8	228.2±35.0	5.0±0.3	39.3±1.7	101.1±14.7	5.7±1.0	38.0±1.7	129.4±19.0
DiVE	29.8±0.2	96.4±0.2	230.4±5.0	26.0±0.7	91.3±0.7	176.8±2.6	13.9±0.2	62.9±1.2	250.0±14.7
LatentShift	8.0±0.6	63.2±3.0	111.8±4.3	11.2±2.2	64.6±7.4	117.7±11.2	8.9±0.2	52.0±1.1	211.9±4.2
D'artagnan	24.2±3.6	80.0±3.1	266.0±15.9	8.8±1.2	56.6±2.5	108.4±26.4	14.2±2.4	61.3±4.2	248.1±8.3
DVCE	1.3±0.2	11.9±2.1	46.7±3.6	1.3±0.2	11.9±2.1	46.7±3.6	1.6±0.1	10.1±0.5	67.1±4.6

TABLE IV: Fidelity of functional connectivity (FC) features of counterfactual samples for HCP-Rest, HCP-Task and ID1000 based on the MMP atlas.

	HCP-Rest			HCP-Task			ID1000		
	Prox. ↓	Spar. ↓	FID ↓	Prox. ↓	Spar. ↓	FID ↓	Prox. ↓	Spar. ↓	FID ↓
DreaMR	**0.7±0.1**	**3.5±0.6**	**26.9±1.5**	**1.0±0.0**	**5.0±0.4**	**6.6±0.5**	**0.9±0.0**	**1.6±0.0**	**24.0±0.9**
DiME	0.9±0.1	6.7±1.6	39.5±1.7	1.8±0.2	13.6±2.4	15.8±1.7	1.6±0.0	6.3±0.2	46.7±0.8
DiffSCM	3.0±0.2	25.1±1.2	99.2±4.7	3.9±0.2	23.2±1.0	36.6±1.0	2.5±0.3	14.7±1.2	79.7±7.0
CAG	3.4±1.6	29.3±10.4	30.8±2.4	3.2±0.3	25.2±2.2	20.6±2.3	3.7±0.1	20.8±0.2	32.7±3.3
CheXplain	9.8±1.1	58.0±2.8	193.6±17.6	5.5±0.5	38.3±2.6	90.3±0.4	5.3±0.6	30.1±2.4	156.5±22.3
DiVE	25.3±1.1	93.7±1.1	241.5±3.2	23.6±0.5	89.9±0.6	180.2±5.1	16.5±0.4	63.2±0.9	264.7±1.7
LatentShift	6.5±0.4	57.2±3.0	136.9±5.9	10.7±3.8	60.0±14.0	115.0±50.4	8.1±0.3	43.5±1.9	211.3±6.2
D'artagnan	12.8±2.8	69.8±6.7	251.5±15.7	4.6±1.3	35.5±6.7	93.8±36.0	9.4±2.9	46.2±7.9	201.4±29.9
DVCE	1.5±0.2	15.0±3.7	43.9±2.9	2.5±0.3	21.6±3.1	22.5±2.9	1.5±0.1	5.8±0.2	44.2±2.2

TABLE V: Fidelity of spatiotemporal and FC features of counterfactual samples for the HCP-Task dataset. Results based on an aggregate atlas combining ROIs from the MMP and Talairach atlases.

	Spatiotemporal			FC		
	Prox. ↓	Spar. ↓	FID ↓	Prox. ↓	Spar. ↓	FID ↓
DreaMR	**56.5±0.4**	17.6±0.2	**4.9±0.1**	**1.0±0.0**	**5.1±0.2**	**6.1±0.4**
DiME	69.4±0.7	21.7±0.2	9.6±1.3	1.5±0.1	10.5±1.1	14.1±1.1
DiffSCM	76.0±9.0	**14.6±0.9**	21.5±1.5	5.1±0.5	27.0±1.7	55.6±7.5
CAG	155.0±41.9	39.6±5.4	13.5±0.9	2.7±0.2	20.8±1.2	22.5±3.6
CheXplain	202.6±8.6	47.2±8.6	98.2±4.0	4.2±0.1	31.1±0.6	92.8±10.2
DiVE	107.1±0.9	32.7±0.2	129.1±5.9	23.6±1.2	88.2±1.3	182.4±6.0
LatentShift	122.2±55.7	31.3±9.3	123.6±14.5	11.5±2.2	58.5±7.8	141.4±15.1
D'artagnan	64.3±6.9	19.9±2.6	29.6±6.0	5.4±0.3	41.7±1.6	81.5±13.2
DVCE	57.9±0.8	17.8±0.3	19.9±2.1	2.4±0.2	18.6±2.4	26.1±3.8

V. Results

A. Fidelity of Counterfactual fMRI Samples

We first demonstrated DreaMR in explanation of downstream transformer-based classifiers for sex on HCP-Rest and ID1000, and for cognitive task on HCP-Task. Separate analyses were conducted using the Schaefer and MMP atlases to rule out potential concerns regarding bias in ROI definitions. Fig. 3 depicts representative explanation maps for spatiotemporal features of an fMRI sample from the motor task, produced by competing counterfactual as well as attribution [43] and perturbation [69], [70] methods. Attribution methods tend to yield broad-spread maps with poor specificity for local features in the fMRI sample. Perturbation methods can improve capture for some local features, but their results are inconsistent across the fMRI sample given limited sensitivity for global context. Meanwhile, counterfactual methods based on VAE, GAN priors tend to emphasize non-salient features for the motor task, suggesting suboptimal fidelity in counterfactual samples. In contrast to competing methods, DreaMR generates an explanation map that is more closely aligned with prominent spatiotemporal features in the average sample for the motor task, without signs of over-broadening or incoherence. These qualitative assessments suggest that DreaMR can offer high fidelity in counterfactual explanation.

To provide specific explanations, counterfactual generation aims to identify minimal changes in the features of an original fMRI sample that are sufficient to flip the respective classifier decision. Thus, under the condition of successful decision flipping, it is desirable that counterfactual and original samples follow similar data distributions. Based on this notion, the fidelity of counterfactual samples was quantitatively evaluated via proximity, sparsity and FID metrics (see Section IV-C for definitions), which reflect the similarity between the counterfactual and original data distributions. Proximity, sparsity and FID values for spatiotemporal features of counterfactual samples are listed in Table I based on the Schaefer atlas, and in Table II based on the MMP atlas. Note that, in these quantitative evaluations, each competing method yielded successful counterfactuals that were able to flip the classifier decisions for all original samples in the test sets. In general, the relative performance levels of competing methods show similar trends for analyses based on Schaefer versus MMP atlases, suggesting that our results are not unduly biased by ROI definitions. Overall, DreaMR achieves the lowest proximity, sparsity and FID across datasets and atlases ($p<0.05$), except for DiffSCM that yields moderately lower sparsity on HCP-Task and ID1000 datasets, and CAG that yields moderately lower FID on HCP-Rest and on ID1000 with the MMP atlas. Note that, in these exception cases, DreaMR still performs competitively with DiffSCM or CAG, and yields the second-best performance among competing methods. On average across atlases and cross-validation folds, DreaMR outperforms competing methods in proximity by 51.6, sparsity by 13.9, FID by 57.6 on HCP-Rest; proximity by 49.7, sparsity by 11.9, FID by 47.8 on HCP-Task; and proximity by 70.4, sparsity by 16.7, FID by 57.8 on ID1000. These results indicate that DreaMR generates high-fidelity counterfactuals minimally different from the original fMRI samples and closely aligned with the original data distribution in terms of spatiotemporal features.

In the neuroimaging literature, another pervasively analyzed attribute of fMRI data are functional connectivity (FC) features that are commonly associated with cognition-related variables [3]. Thus, we also examined the fidelity of FC features derived from counterfactual fMRI samples. To do this, a counterfactual sample was generated via competing methods on each original fMRI sample. For both counterfactual and original samples, the FC feature between a given pair of brain regions was taken as Pearson's correlation coefficient between respective fMRI time courses [8]. Note that if a particular method yields counterfactual samples that are more similar to the original fMRI samples, then the FC features derived from those counterfactuals are likely to be more similar to the FC features of original samples. Thus, in terms of relative performance among competing methods, results for FC features are expected to follow partly similar patterns with those for spatiotemporal features. Proximity, sparsity and FID values for FC features are listed in Table III based on the Schaefer atlas, and in Table IV based on the MMP atlas. The relative performance levels of competing methods are generally similar based on Schaefer versus MMP atlases, corroborating that our results are not unduly biased by ROI definitions. Overall, DreaMR achieves the top performance in all datasets and atlases ($p<0.05$). On average across atlases and folds, DreaMR outperforms competing methods in proximity by 8.6, sparsity by 42.9, FID by 99.7 on HCP-Rest; proximity by 6.5, sparsity by 35.5, FID by 68.4 on HCP-Task; and proximity by 5.4, sparsity by 29.6, FID by 103.6 on ID1000. These results indicate that DreaMR generates reliable counterfactual samples whose FC features are closely aligned to the distribution of FC features for original fMRI samples, similar to the results on spatiotemporal features. Note, however, that FC features reflect time-aggregated connectivity measures that typically follow a lower dimensional distribution compared to spatiotemporal features, and the distributional attributes of the two feature sets show notable differences [3]. As such, a simple comparison of fidelity metrics for FC versus spatiotemporal features might yield misleading impressions regarding relative success in counterfactual explanation, and further work is warranted to elucidate this challenging question.

Certain cognitive tasks might involve brain regions not only within but also outside the cerebral cortex, for which the Schaefer and MMP atlases provide a tessellation [71]. For instance, it has been suggested that motor and working memory tasks evoke responses in cerebellar regions, whereas gambling tasks can evoke responses in subcortical regions [57]. To provide a broader assessment across the brain, we conducted a separate analysis based on an aggregate atlas that combined ROI definitions for cortical regions from the MMP atlas with those for cerebellar and subcortical regions from the Talairach atlas. A downstream transformer-based classifier was first trained to detect cognitive tasks on the HCP-Task dataset, and competing methods were then employed for counterfactual explanation. Proximity, sparsity and FID values for both spatiotemporal and FC features of counterfactual samples are listed in Table V. For both types of features, we find that the fidelity metrics of competing methods show very similar distributions based on MMP versus aggregate atlases, where DreaMR generally yields the lowest metric values among competing methods. These results suggest that DreaMR generates high-quality counterfactuals that are closely aligned with the distribution of original fMRI samples not only in cortical but also in cerebellar and subcortical regions.

TABLE VI: Detection performance of a linear classifier trained using counterfactual samples was measured via accuracy (Acc.) and F1 metrics. Results listed for HCP-Rest, HCP-Task and ID1000 based on the Schaefer atlas as mean±std across five cross-validation folds.

	HCP-Rest		HCP-Task		ID1000	
	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
DreaMR	**78.0±2.1**	**80.0±2.3**	**92.7±3.0**	**92.8±3.0**	**78.0±2.4**	**75.8±5.2**
DiME	47.8±2.7	63.2±1.6	84.6±2.5	83.8±2.9	68.9±6.4	74.9±4.3
DiffSCM	53.0±1.7	63.0±1.2	88.8±4.4	89.2±3.9	52.6±3.5	53.1±1.1
CAG	62.8±11.6	70.8±5.1	69.0±5.7	69.1±6.4	57.50±8.4	69.8±5.0
CheXplain	52.6±2.0	64.2±0.4	19.6±5.9	14.1±4.7	53.0±3.7	67.1±1.8
DiVE	64.9±10.6	71.1±5.3	64.5±10.9	63.1±11.7	66.1±9.1	74.2±4.8
LatentShift	52.1±4.3	65.6±3.1	76.9±3.1	76.3±3.9	51.5±3.2	66.6±1.8
D'artagnan	55.3±6.6	47.4±17.4	56.5±6.3	53.9±6.4	49.1±2.3	40.0±29.9
DVCE	58.5±7.6	68.5±3.9	76.1±12.8	74.7±14.2	60.5±9.0	71.1±6.8

TABLE VII: Detection performance of a linear classifier trained using counterfactual samples for HCP-Task based on the aggregate atlas. Results listed separately for cortical and non-cortical ROIs.

	Cortical		Non-cortical	
	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
DreaMR	**90.6±2.0**	**90.4±2.0**	**24.9±2.5**	**21.3±3.8**
DiME	83.3±5.7	83.1±5.9	17.3±1.8	12.3±3.2
DiffSCM	68.7±6.8	66.0±6.0	19.8±1.9	12.8±1.9
CAG	66.5±6.7	65.6±7.0	21.5±3.8	15.5±6.0
CheXplain	21.0±5.3	13.7±4.3	14.1±0.5	6.4±1.6
DiVE	82.9±3.0	82.8±3.2	20.9±3.4	12.2±4.3
LatentShift	82.1±2.5	81.9±2.7	19.7±2.0	16.7±2.8
D'artagnan	62.9±8.1	61.8±7.0	16.9±0.7	10.6±3.3
DVCE	19.1±4.3	11.4±5.2	15.0±0.4	5.0±0.8

B. Explanations based on Counterfactual Samples

Neuroscience studies routinely report that individual cognition-related variables (e.g., sex or cognitive task) are associated with characteristic patterns of brain responses [3], [60]. Given a downstream deep-learning classifier trained to detect such variables, the goal of counterfactual explanation is to render important input features explicitly observable in the differences between counterfactual and original fMRI samples, such that classifier decisions can be directly interpreted without the need for further processing. This bears out that the differences between counterfactual and original fMRI samples should ideally follow the differences between characteristic response patterns for the target versus original variables. As such, the explanatory capability of a counterfactual method can be assessed by examining the discriminative information that the counterfactual samples explicitly carry regarding the target versus original variables. Note that non-linear classifiers (e.g., transformer-based classifiers explained here) subject their inputs to multiple levels of processing to unravel hidden features that are not directly observable in

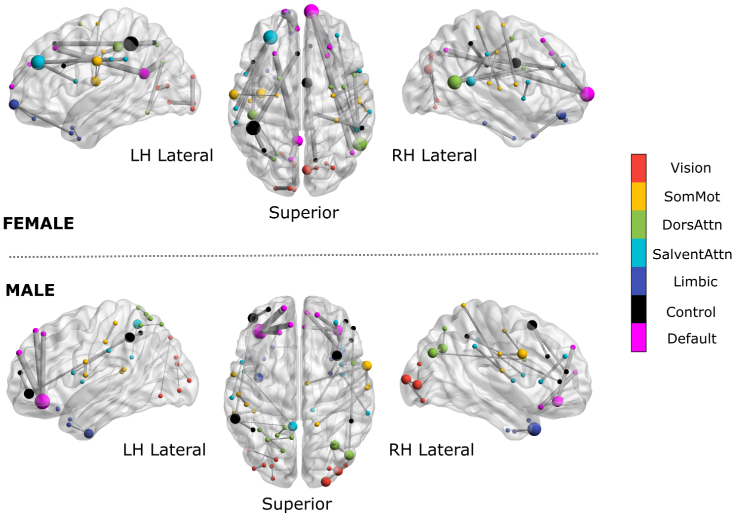

Fig. 4: Cortical explanation maps produced by DreaMR for female (top panel) and male (bottom panel) sexes on the HCP-Rest dataset, averaged across five cross-validation folds. To obtain the cortical maps, counterfactual samples were generated for each original fMRI sample to flip the respective decision of a deep fMRI classifier for sex detection. FC features of original and counterfactual samples were separately derived, and the differences between the two sets of samples were averaged across subjects. Important FC features were determined by selecting the features showing the top 5% of differences. Each important region-of-interest (ROI) in the brain is marked with a dot on the anatomical template, and connectivity between ROIs is shown with a bar. Dot size denotes the importance of the ROI, and bar thickness denotes the importance of the connection. ROIs are colored according to the functional network they belong to (see legend). Lateral and superior brain views are displayed. LH: left hemisphere, RH: right hemisphere; SomMot: somatomotor; DorsAttn: dorsal attention; SalventAttn: Salience/ventral attention.

their inputs, and so their results can reflect spurious bias from the architecture of the classification model rather than the explicit information content of the inputs [3], [72]. To avoid potential biases, we adopted a linear classifier to assess the discriminative information that is explicitly observable in counterfactual samples [73]. A separate linear classifier was fit using the counterfactual samples generated by each competing method [36]. The linear classifiers were then tested on the original fMRI samples to detect the same variables as the nonlinear transformer-based classifiers. Table VI lists accuracy and F1 for linear classifiers based on competing methods and the Schaefer atlas. DreaMR elicits the highest detection performance in all cases ($p<0.05$). On average across folds, DreaMR outperforms competing methods in accuracy by 22.1%, F1 by 15.9% on HCP-Rest; accuracy by 25.7%, F1 by 27.3% on HCP-Task; accuracy by 20.6%, F1 by 11.2% on ID1000. This finding indicates that DreaMR generates counterfactual samples that better capture the characteristic differences in response patterns between distinct cognition-related variables. We also questioned whether this benefit is evident beyond the cerebral cortex. For this purpose, we examined the discriminative information in counterfactual samples for HCP-Task based on the aggregate atlas including non-cortical ROIs. Each counterfactual sample generated based on the aggregate atlas was split into two subsamples, one containing only cortical ROIs and the other containing only non-cortical ROIs. Accuracy and F1 metrics of separate linear classifiers fit for cortical versus non-cortical ROIs are listed in Table VII. Among competing methods, DreaMR attains the highest detection performance for both cortical and non-cortical ROIs ($p<0.05$). Note that while non-cortical ROIs generally show above chance-level detection performance (i.e., $>14.3\%$ accuracy), the detection performance for cortical ROIs is substantially higher. This finding implies that fMRI responses in cortical regions represent a greater amount of information about cognitive tasks included in the HCP-Task dataset than those in non-cortical regions.

Next, counterfactual samples from DreaMR were analyzed to derive cortical explanation maps that reflect the important FC features associated with individual classes detected by downstream transformer-based classifiers. To do this, differences of FC features between original and counterfactual samples were computed. Important features showing the top 5% of differences across subjects were determined. As a representative case, cortical explanation maps for each sex on HCP-Rest are displayed in Fig. 4 based on the Schaefer atlas. For females, relatively important features are observed in dorso-medial prefrontal and medial-posterior segments of the default mode network, parietal and medial segments of the control network, orbitofrontal segments of the limbic network, temporal-occipital and prefrontal segments of the salience/ventral attention network, posterior segments of the dorsal attention network, central segments of the somatamotor network, and medial segments of the vision network. Meanwhile, for males, important features are observed in ventral- and dorsal-prefrontal segments of the default mode network, lateral-parietal and lateral-prefrontal segments of the control network, temporal-pole segments of the limbic network, medial segments of the salience/ventral attention network, posterior segments of the dorsal attention network, precentral segments of the somatamotor network, and lateral segments of the vision network. Note that these patterns in the cortical distribution of important FC features are closely aligned with the neuroimaging literature that identifies prominent sex-related differences in resting-state connectivity across the default mode, control, limbic, attention and somatomotor networks [10], [74]. In particular, ROIs distributed across relatively medial segments of default mode and control networks, associated with social cognition and executive control, have been suggested to show relatively higher FC values in females than males [75]. Similarly, ROIs in relatively orbitofrontal segments of the limbic network associated with emotion and memory have been reported to show higher FC values in females than males [76], [77]. While both sexes can show important FC features in the somatosensory network associated with motor tasks and in attention networks associated with multi-sensory spatial tasks, FC values across these networks have been reported to show stronger tendency for left lateralization in females and right lateralization in males [75], [78], as also evident in our cortical explanation maps.

TABLE VIII: Inference times (Inf., msec) and memory load (Mem., gigabytes) per generation of a counterfactual fMRI sample.

	HCP-Rest		HCP-Task		ID1000	
	Inf.	Mem.	Inf.	Mem.	Inf.	Mem.
DreaMR	2735	3.7	1173	2.7	1114	2.7
DiME	22104	6.0	14806	4.2	14918	4.2
DiffSCM	14327	3.5	7049	2.3	7343	2.4
CAG	116	4.3	45	3.2	45	3.3
CheXplain	384	3.0	491	2.9	282	2.9
DiVE	1750	3.8	1121	3.1	1149	2.8
LatentShift	967	3.7	850	2.9	829	2.2
D'artagnan	146	3.3	62	3.3	59	3.3
DVCE	23456	5.1	10740	4.1	10293	4.1

TABLE IX: Fidelity of spatiotemporal features of counterfactual samples generated by DreaMR variants formed without a transformer, without Langevin sampling, without fractional diffusion, without multi-phase distillation, and with guidance from classifiers trained on noisy fMRI samples.

	Prox. ↓	Spar. ↓	FID ↓
DreaMR	**41.4±0.5**	**11.7±0.2**	21.2±0.8
w/o transformer	48.2±0.6	14.5±0.2	23.8±1.0
w/o Langevin	43.9±0.2	12.8±0.1	**20.5±1.0**
w/o fraction	42.5±1.0	12.2±0.4	21.1±1.1
w/o mp distillation	47.9±0.7	14.5±0.3	22.1±1.3
w noisy-classifier	41.6±0.6	11.8±0.2	21.3±1.1

The overall consistency of our findings with the neuroimaging literature suggests that DreaMR is a promising framework to identify fMRI features that are characteristic to individual variables such as sex that influence cognitive function, and hence drive the decisions of downstream fMRI classifiers for these variables.

C. Computational Efficiency

A practical concern for counterfactual generation is the efficiency in resampling of original fMRI samples. Table VIII lists inference time and memory use of competing methods based on the Schaefer atlas. In inference time, non-iterative CAG and D'artagnan methods are the fastest, whereas iterative DiME, DiffSCM and DVCE based on conventional diffusion priors are notably slower. The remaining iterative methods including DreaMR require intermediate inference times between the two extremes. Note that the FMD prior in DreaMR enables substantially improved efficiency over conventional diffusion priors, and comparable efficiency with VAE and GAN priors. In terms of memory load, DreaMR has comparable demand to methods based on VAE and GAN priors, comparable demand to DiffSCM, and moderately lower demand than DiME.

D. Ablation Studies

A series of ablation studies were conducted on the HCP-Rest dataset to assess the importance of the main design elements in DreaMR. For this purpose, several ablated variants of DreaMR were considered. First, we examined the influence of utilizing a transformer architecture, calculating the denoised sample estimate for classifier guidance based on iterated Langevin sampling, using fractional diffusion, using multi-phase distillation, and using classifier gradients on denoised sample estimates. A 'w/o transformer' variant replaced the transformer with the common UNet architecture for diffusion priors [28]. A 'w/o Langevin' variant replaced the denoised sample estimate obtained through iterated Langevin sampling across diffusion fractions with a single-shot estimate predicted by the denoising network of the current fraction. A 'w/o fraction' variant used a single fraction for the diffusion prior. A 'w/o mp distillation' variant used a single-phase distilled diffusion prior. A 'w noisy-classifier' variant computed the conditional score function for classifier guidance by training a separate classifier on noisy fMRI samples instead of using the original classifier on clean samples. Proximity, sparsity and FID values for spatiotemporal features of counterfactual samples are listed in Table IX based on the Schaefer atlas. Overall, we find that DreaMR outperforms all ablated variants, except for 'w/o Langevin' that yields moderately lower FID and 'w noisy-classifier' that generally performs similarly. Compared to DreaMR, 'w/o transformer' yields 6.8 higher proximity (16.4% performance loss), 2.8 higher sparsity (23.9% loss), 2.6 higher FID (12.3% loss); 'w/o Langevin' yields 2.5 higher proximity (6.0% loss), 1.1 higher sparsity (9.4% loss); 'w/o fraction' yields 1.1 higher proximity (2.7% loss), 0.5 higher sparsity (4.3% loss), and 0.1 higher FID (0.5% loss); 'w/o mp distillation' yields 6.5 higher proximity (15.7% loss), 2.8 higher sparsity (23.9% loss), 0.9 higher FID (4.2% loss). While diffusion-based elements elicit relatively limited benefits for FID, their contributions to proximity and sparsity are more comparable to the transformer architecture. Note that, in calculating the score function for classifier guidance, the 'w noisy-classifier' variant replaces the surrogate gradients based on the original classifier with the true gradients of a classifier separately trained on noisy fMRI samples. We observe that DreaMR and 'w noisy-classifier' perform very similarly, suggesting that the gradients of the original classifier serve as a successful surrogate. Taken together, these results suggest that each interrogated design element contributes significantly to method performance.

TABLE X: Fidelity of spatiotemporal features of counterfactual samples for DreaMR variants formed by implementing the FMD prior with varying number of fractions (F) and number of distillation phases (P).

Fractions				Distillation phases			
F	Prox.↓	Spar. ↓	FID ↓	P	Prox.↓	Spar. ↓	FID ↓
1	42.5±1.0	12.2±0.4	21.1±1.1	5	45.1±0.5	13.3±0.2	**19.7±1.1**
2	41.3±0.7	11.7±0.3	**21.0±0.8**	6	43.6±0.5	12.7±0.2	20.1±0.6
4	41.4±0.5	11.7±0.2	21.2±0.8	7	**41.4±0.5**	**11.7±0.2**	21.2±0.8
8	**41.0±0.4**	**11.6±0.2**	21.8±0.5	8	44.9±0.6	13.2±0.3	**19.7±1.1**

We then examined the influence of the number of diffusion fractions (F) and distillation phases (P) on the fidelity of the counterfactual samples generated by the FMD prior. To do this, variants of DreaMR were trained for varying F while P=7, and for varying P while F=4. Table X lists proximity, sparsity and FID values for spatiotemporal features of counterfactual samples produced by DreaMR variants based on the Schaefer atlas. We find that the selected values of F=4, P=7 generally attain near-optimal performance. While F=8 yields slightly lower proximity and sparsity values, F=4 is preferable as it offers higher training efficiency by halving the number of distinct denoising networks to be learned. While P=5 and $P = 8$ yield slightly lower FID, they have moderately higher proximity and sparsity values, and P=7 offers over sixteen times faster sampling than $P = 5$ since inference time for counterfactual generation scales nearly quadratically with the number of diffusion steps.

VI. Discussion

To our knowledge, DreaMR is the first diffusion-based counterfactual explanation method for fMRI analysis, and it is the first method in the literature that synergistically combines multi-phase distillation with fractional diffusion based on an efficient transformer backbone. This unique design enables DreaMR to simultaneously attain high fidelity and efficiency in counterfactual generation, unlike conventional diffusion priors that face a characteristic fidelity-efficiency trade-off. Ablation studies demonstrate that each design element contributes significantly to explanation performance. Comparison studies demonstrate that DreaMR provides more reliable explanations than competing methods based on state-of-the-art VAE, GAN and diffusion priors. Note that the diffusion baselines DiffSCM, DiVE and DVCE are devoid of the unique technical innovations that DreaMR embodies. As such, our results collectively indicate that DreaMR helps push the performance envelope in counterfactual explanation. At the same time, DreaMR achieves significantly improved efficiency in sample generation over conventional diffusion priors, resulting in brief inference times that approach the repetition time (TR) of common fMRI acquisitions.

The inference efficiency of DreaMR can be beneficial in real-time fMRI applications such as intraoperative planning, brain-computer interfaces, communication with locked-in patients, and therapeutic regulation of brain activations [29]. Real-time fMRI requires rapid detection of cognition-related variables to provide timely feedback for user intervention (e.g., a surgeon deciding on resection margins while the subject performs a cognitive task, or a patient trying to control brain activity to suppress tremor). Deep-learning classifiers promise sensitive detection of cognitive variables from real-time fMRI data, albeit poor interpretability can limit user trust in classifier decisions [79]. With its fast inference, DreaMR can facilitate rapid intervention by avoiding undesired delays in counterfactual explanation. Another domain that can benefit from the efficiency of DreaMR is cohort studies that involve large-scale analyses of fMRI scans from a subject population to associate features in imaging data with specific classes of neurological disease [30]. Although deep-learning classifiers have emerged as state-of-the-art tools in establishing such associations, they still suffer from limited interpretability [30]. Note that explaining deep-learning classifiers with conventional diffusion priors can be computationally burdening for fMRI datasets comprising thousands of subjects, tens of different disease classes, and various different candidates for classifier architectures. In this context, adopting DreaMR for explaining classifier decisions can help improve reliability and efficiency of data analyses in cohort fMRI studies.

Recent neuroimaging studies employing deep-learning classifiers have reported promising results on detection of prevalent neurodegenerative (e.g., mild cognitive impairment, Alzheimer's) and neurodevelopmental (e.g., attention deficit hyperactivity disorder, Autism) conditions from resting-state fMRI scans [7], [8]. As these fMRI studies are increasingly adopting complex network architectures such as transformers [35], providing reliable explanations for classifier-driven diagnostic decisions via DreaMR can offer important benefits in terms of building user trust and accelerating translation to clinical use. The utility of fMRI classifiers in clinical diagnostics inevitably depends on the prospect of fMRI for capturing disease-related signatures in the nervous system, and current evidence suggests that fMRI can be a valuable component of multi-modal disease assessments [80], [81]. While we primarily demonstrated DreaMR on fMRI scans in the current study, it is important to note that the proposed explanation method can also be adopted to explain classifiers built on other types of imaging data including structural or dynamic MRI scans that are pervasive in diagnosis of musculoskeletal, neurological and cardiovascular diseases.

Limitations

Several lines of technical limitations could be addressed to further improve DreaMR. Following common practice, here we analyzed fMRI data preprocessed to register brain volumes onto an anatomical template. This procedure facilitates comprehensive and consistent region definitions across subjects based on an atlas [1], yet registration onto a template can yield spatial information losses. Potential losses might be mitigated by backprojecting atlas-based region definitions onto the brain spaces of individual subjects [82], and preserving spatial representations of individual subjects via enhanced localization mechanisms in transformer models [83], [84].

The diffusion priors in DreaMR were trained here from scratch on public fMRI datasets comprising several hundred subjects. Literature suggests that such sizable datasets might be critical in adequate training of diffusion priors. In resource-limited application domains where data are scarce, fine tuning diffusion priors pretrained on time-series data such as audio waveforms might help improve learning [85]. Knowledge distillation from adversarial priors might also be employed to facilitate training of diffusion priors [86], [87].

Counterfactual generation with a diffusion prior requires formation of an initial sample that reflects a noisy representation of the original data sample. In this study, initial samples were derived by adding random Gaussian noise onto the original samples, following the forward diffusion process. Several other techniques have been proposed in the literature to obtain the initial noisy samples [50], [88]. As these techniques can alter the distribution of initial noisy samples, they might also influence the fidelity of resultant counterfactual samples. For instance, it has been suggested that DDIM inversion improves distributional similarity between counterfactual and original samples, as it derives noisy representations based on estimates of noise components in original samples [50]. Yet, such inversion can also elevate nuisance noise correlations between counterfactual and original samples across the ROI and temporal dimensions, potentially restricting the representational capacity of the diffusion prior. In early phases of the study, we implemented a variant of DreaMR based on DDIM inversion, and while this variant yielded moderate benefits in fidelity of spatiotemporal features, it performed suboptimally in fidelity of FC features and in linear classification analyses that reflect the level of discriminative information about cognitive variables captured by counterfactual samples (i.e., for the representative sex-detection task on HCP-Rest, 78.0% accuracy, 80.0% F1 with Gaussian noise addition versus 62.9% accuracy, 55.4% F1 with DDIM inversion). That said, further work is warranted to comprehensively evaluate the relative benefits of Gaussian noise addition versus alternative techniques to obtain noisy representations in the context of counterfactual explanations of deep fMRI classifiers.

Here, DreaMR was implemented with uniform diffusion fractions of equal duration to give similar emphasis on each fraction. Alternatively, non-uniform fractions could be employed to help account for potential variability in the denoising task across the diffusion process due to varying noise levels and feature details. When a sufficiently large number of fractions are prescribed that help effectively capture task variability, we would not expect substantial differences between uniform versus non-uniform fractions. Yet, when a relatively limited number of fractions are prescribed, non-uniform fractions might offer potential performance benefits. In such cases, longer-duration fractions can be employed in diffusion segments of relatively slow variation, and shorter-duration fractions can be employed in segments of relatively fast variation in the characteristics of the denoising task. Future work is warranted to systematically assess the performance benefits of non-uniform versus uniform diffusion fractions.

To shorten sampling times for counterfactual generation, DreaMR leverages multi-phase distillation of its fractional diffusion prior. This distillation procedure lowers computational burden for the testing stage, at the expense of additional burden introduced by distillation in the training stage. Corroborating recent reports, here we observed that distillation via a moderate number of phases helps improve sample quality compared to using a relatively limited number of phases [54]. Still, when needed, single-phase distillation procedures can be adopted to improve efficiency during the training stage at the expense of moderate losses in sample quality. Here, we obtained efficient explanations for fMRI samples lasting hundreds of time frames. Naturally, the burden of counterfactual generation during the testing phase grows with the temporal dimension of fMRI scans. In applications where the load becomes excessive due to high temporal resolution or long scan duration, the FMD prior might be combined with other efficient sampling approaches for accelerated diffusion [65], [89]. During the testing stage, DreaMR uses a nested algorithm where an inner loop produces estimates for the clean sample via Langevin sampling, and this estimate is then used to compute classifier guidance that drives the outer loop to produce the counterfactual sample. This algorithm was adopted as the inner loop does not induce unduly burden given the large step sizes prescribed for DreaMR, and since it was observed to yield higher sample quality. That said, the efficiency and simplicity of the sampling algorithm might be improved by discarding the inner loop and instead using the single-shot network estimate for the clean sample. Further work is warranted to systematically assess the benefits of various sampling algorithms in terms of sample fidelity versus efficiency during counterfactual generation.

Future Work

Here, we utilized DreaMR to explain transformer-based classifiers that predict discrete cognitive states given brain responses. Several extensions can be pursued to expand the scope of the proposed methodology. First, classifiers based on alternative convolutional or recurrent architectures might be considered [6], [10], [64]. Since counterfactual generation is a model-agnostic framework, DreaMR can in principle be adopted to other architectures without modification. Second, many neurodegenerative and neurodevelopmental diseases are reported to have complementary biomarkers in fMRI as well as anatomical or diffusion-weighted MRI [80], [81]. A more performant classifier for disease-related cognitive states could be attained by using multi-modal images as input. DreaMR can be adopted for explaining such multi-modal classifiers by training a multi-modal FMD prior.

DreaMR also holds potential for explaining regression models in fMRI analysis based on deep neural networks. The human brain represents information on continuous stimulus or task variables, which can be decoded from brain responses via regression [60]. Deep regression models can also be employed to predict voxel- or ROI-wise responses given stimulus variables [90]. To adopt DreaMR, guidance from classification loss during counterfactual generation could be replaced with guidance from regression losses. It remains important future work to assess the efficacy of DreaMR in a broader set of explanation tasks in fMRI analysis.

It might be possible to adopt DreaMR to explain deep-learning models used to analyze other types of neuroimaging data beyond fMRI scans. Neuroimaging studies principally record spatiotemporal features of neural activity while cognitive variables are experimentally manipulated. For instance, electroencephalography measures local field potentials [91] and magnetoencephalography measures local magnetic fields caused by neural activity [92], near-infrared spectroscopy measures hemodynamic changes consequent to neural activity [93], and microelectrode arrays directly measure multi-unit activity [94]. Inferences are then drawn by analyzing the association between measured features and cognitive variables. In this context, deep-learning classifiers have recently gained traction as a leading approach to detect cognitive variables from measured features [95]. Yet, their limited interpretability hampers trust in inferences based on classification performance. In principle, DreaMR could help interrogate the specific spatiotemporal features of neural activity that contribute to classifier decisions, thereby enhancing the utility of classification analyses in neuroimaging studies.

VII. Conclusion

In this study, we introduced a novel counterfactual explanation method for fMRI based on a fractional multi-phase-distilled diffusion prior. Demonstrations on resting-state and task-based fMRI indicate that DreaMR achieves higher sampling efficiency and fidelity against competing counterfactual methods, facilitating interpretation of downstream classifier decisions. Therefore, the proposed method holds great promise in enabling explainable analysis of multi-variate fMRI data with deep-learning models.

References

[1] M. J. Singleton, "Functional magnetic resonance imaging," *Yale J Biol Med*, vol. 82, no. 4, p. 233, 2009.

[2] S. Nishimoto *et al.*, "Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies," *Curr Biol*, vol. 21, no. 19, pp. 1641–1646, 2011.

[3] F. Pereira *et al.*, "Machine learning classifiers and fMRI: a tutorial overview," *NeuroImage*, vol. 45, no. 1, pp. S199–S209, 2009.

[4] H. Jang *et al.*, "Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks," *NeuroImage*, vol. 145, pp. 314–328, 2017.

[5] H. Huang *et al.*, "Modeling task fMRI data via deep convolutional autoencoder," *IEEE Trans Med Imaging*, vol. 37, no. 7, pp. 1551–1561, 2017.

[6] J. Kawahara *et al.*, "BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment," *NeuroImage*, vol. 146, pp. 1038–1049, 2017.

[7] S. Parisot *et al.*, "Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease," *Med Image Anal*, vol. 48, pp. 117–130, 2018.

[8] T.-E. Kam *et al.*, "Deep learning of static and dynamic brain functional networks for early MCI detection," *IEEE Trans Med Imaging*, vol. 39, no. 2, pp. 478–487, 2019.

[9] Y. Li *et al.*, "Deep spatial-temporal feature fusion from adaptive dynamic functional connectivity for MCI identification," *IEEE Trans Med Imaging*, vol. 39, no. 9, pp. 2818–2830, 2020.

[10] B.-H. Kim *et al.*, "Learning Dynamic Graph Representation of Brain Connectome with Spatio-Temporal Attention," in *NeurIPS*, no. 330. Curran Associates Inc., 2021, pp. 4314–4327.

[11] L. Wang *et al.*, "Graph convolutional network for fMRI analysis based on connectivity neighborhood," *Net Neurosci*, vol. 5, no. 1, pp. 83–95, 2021.

[12] X. Wang *et al.*, "Contrastive functional connectivity graph learning for population-based fmri classification," in *MICCAI, Lect Notes Comput Sci*, vol. 13431. Springer, 2022, pp. 221–230.

[13] B. H. van der Velden *et al.*, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Med Image Anal*, vol. 79, p. 102470, 2022.

[14] T. Matsui *et al.*, "Counterfactual explanation of brain activity classifiers using image-to-image transfer by generative adversarial network," *Front Neuroinf*, vol. 15, p. 79, 2022.

[15] P. Rodriguez *et al.*, "Beyond trivial counterfactual explanations with diverse valuable explanations," in *IEEE ICCV*, 2021, pp. 1056–1065.

[16] B.-H. Kim *et al.*, "Understanding graph isomorphism network for rs-fMRI functional connectivity analysis," *Front Neurosci*, p. 630, 2020.

[17] L. Tomaz Da Silva *et al.*, "Visual Explanation for Identification of the Brain Bases for Developmental Dyslexia on fMRI Data," *Front Comput Neurosci*, vol. 15, p. 594659, 2021.

[18] Y. Kazemi *et al.*, "A deep learning pipeline to classify different stages of Alzheimer's disease from fMRI data," in *IEEE CIBCB*, 2018, pp. 1–8.

[19] B. J. Devereux *et al.*, "Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway," *Sci Rep*, vol. 8, no. 1, p. 10636, 2018.

[20] J. P. Cohen *et al.*, "Gifsplanation via Latent Shift: A Simple Autoencoder Approach to Counterfactual Generation for Chest X-rays," *arXiv:2102.09475*, 2021.

[21] Y. Korkmaz *et al.*, "Unsupervised MRI reconstruction via zero-shot learned adversarial transformers," *IEEE Trans Med Imaging*, vol. 41, no. 7, pp. 1747–1763, 2022.

[22] J. Ho *et al.*, "Denoising diffusion probabilistic models," in *NeurIPS*, no. 574. Curran Associates Inc., 2020, pp. 6840–6851.

[23] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," *arXiv:2011.13456*, 2020.

[24] P. Sanchez *et al.*, "What is healthy? generative counterfactual diffusion for lesion localization," in *DGM4MICCAI, Lect Notes Comput Sci*, vol. 13609. Springer, 2022, pp. 34–44.

[25] J. Wolleb *et al.*, "Diffusion models for medical anomaly detection," in *MICCAI, Lect Notes Comput Sci*, vol. 13438. Springer, 2022, pp. 35–45.

[26] W. H. L. Pinaya *et al.*, "Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models," *arXiv:2206.03461*, 2022.

[27] C. I. Bercea *et al.*, "Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models," *arXiv:2305.19643*, 2023.

[28] J. Song *et al.*, "Denoising diffusion implicit models," *arXiv:2010.02502*, 2020.

[29] R. C. deCharms, "Applications of real-time fMRI." *Nat Rev Neurosci*, vol. 9, no. 9, pp. 720–729, 2008.

[30] S. Vieira *et al.*, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neurosci Biobehav Rev*, vol. 74, pp. 58–75, 2017.

[31] S. I. Ktena *et al.*, "Metric learning with spectral graph convolutions on brain connectivity networks," *NeuroImage*, vol. 169, pp. 431–442, 2018.

[32] I. Sivgin *et al.*, "A plug-in graph neural network to boost temporal sensitivity in fMRI analysis," *IEEE J Biomed Health Inf*, vol. 28, no. 9, pp. 5323–5334, 2024.

[33] I. Malkiel *et al.*, "Pre-training and Fine-tuning Transformers for fMRI Prediction Tasks," *arXiv:2112.05761*, 2021.

[34] X. Yu *et al.*, "Disentangling Spatial-Temporal Functional Brain Networks via Twin-Transformers," *arXiv:2204.09225*, 2022.

[35] J. Zhang *et al.*, "Diffusion Kernel Attention Network for Brain Disorder Classification," *IEEE Trans Med Imaging*, vol. 41, no. 10, pp. 2814–2827, 2022.

[36] H. A. Bedel *et al.*, "BolT: Fused window transformers for fMRI time series analysis," *Med Image Anal*, vol. 88, p. 102841, 2023.

[37] X. Li *et al.*, "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Med Image Anal*, vol. 65, p. 101765, 2020.

[38] A. Riaz *et al.*, "DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI," *J Neurosci Met*, vol. 335, p. 108506, 2020.

[39] C. A. Ellis *et al.*, "An Approach for Estimating Explanation Uncertainty in fMRI dFNC Classification," in *IEEE BIBE*, 2022, pp. 297–300.

[40] A. Gotsopoulos *et al.*, "Reproducibility of importance extraction methods in neural network based fMRI classification," *NeuroImage*, vol. 181, pp. 44–54, 2018.

[41] H. Vu *et al.*, "fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations," *NeuroImage*, vol. 223, p. 117328, 2020.

[42] S. Arslan *et al.*, "Graph saliency maps through spectral convolutional networks: Application to sex classification with brain connectivity," in *MICCAI GRAIL-MIC Workshop, Lect Notes Comput Sci*. Springer, 2018, pp. 3–13.

[43] Q.-H. Lin *et al.*, "SSPNet: An interpretable 3D-CNN for classification of schizophrenia using phase maps of resting-state complex-valued fMRI data," *Med Image Anal*, vol. 79, p. 102430, 2022.

[44] Y. Li *et al.*, "Hypernetwork construction and feature fusion analysis based on sparse group lasso method on fMRI dataset," *Front Neurosci*, vol. 14, p. 60, 2020.

[45] J. Hu *et al.*, "Interpretable learning approaches in resting-state functional connectivity analysis: the case of autism spectrum disorder," *Comput Math Methods Med*, p. 1394830, 2020.

[46] P. M. N. Dos Santos *et al.*, "Assessing atypical brain functional connectivity development: An approach based on generative adversarial networks," *Front Neurosci*, vol. 16, p. 1025492, 2023.

[47] H. Reynaud *et al.*, "D'artagnan: Counterfactual video generation," in *MICCAI, Lect Notes Comput Sci*, vol. 13438. Springer, 2022, pp. 599–609.

[48] N. Pawlowski *et al.*, "Deep structural causal models for tractable counterfactual inference," in *NeurIPS*, no. 73. Curran Associates Inc., 2020, pp. 857–869.

[49] M. Augustin *et al.*, "Diffusion visual counterfactual explanations," in *NeurIPS*, no. 27. Curran Associates Inc., 2022, pp. 364–377.

[50] P. Sanchez *et al.*, "Diffusion causal models for counterfactual estimation," *arXiv:2202.10166*, 2022.

[51] G. Jeanneret *et al.*, "Diffusion models for counterfactual explanations," in *ACCV, Lect Notes Comput Sci*, vol. 13847. Springer, 2022, pp. 219–237.

[52] R. Liégeois *et al.*, "Resting brain dynamics at different timescales capture distinct aspects of human behavior," *Nat Comm*, vol. 10, no. 1, pp. 1–9, 2019.

[53] H. Chung *et al.*, "Improving diffusion models for inverse problems using manifold constraints," in *NeurIPS*, vol. 35. Curran Associates Inc., 2022, pp. 25 683–25 696.

[54] T. Salimans *et al.*, "Progressive distillation for fast sampling of diffusion models," *arXiv:2202.00512*, 2022.

[55] W. Luo, "A comprehensive survey on knowledge distillation of diffusion models," *arXiv:2304.04262*, 2023.

[56] Y. Balaji *et al.*, "ediffi: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv:2211.01324*, 2022.

[57] D. C. Van Essen *et al.*, "The WU-Minn human connectome project: an overview," *NeuroImage*, vol. 80, pp. 62–79, 2013.

[58] L. Snoek *et al.*, "The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses," *Sci Data*, vol. 8, no. 1, pp. 1–23, 2021.

[59] M. F. Glasser *et al.*, "The minimal preprocessing pipelines for the human connectome project," *NeuroImage*, vol. 80, pp. 105–124, 2013.

[60] T. Çukur *et al.*, "Attention during natural vision warps semantic representation across the human brain," *Nat Neurosci*, vol. 16, no. 6, pp. 763–770, 2013.

[61] A. Schaefer *et al.*, "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI," *Cereb Cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.

[62] M. F. Glasser *et al.*, "A multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, 2016.

[63] J. L. Lancaster *et al.*, "Automated talairach atlas labels for functional brain mapping," *Hum Brain Map*, vol. 10, no. 3, pp. 120–131, 2000.

[64] X. Li *et al.*, "BrainGNN: Interpretable brain graph neural network for fMRI analysis," *Med Image Anal*, vol. 74, p. 102233, 2021.

[65] A. Güngör *et al.*, "Adaptive diffusion priors for accelerated MRI reconstruction," *Med Image Anal*, vol. 88, p. 102872, 2023.

[66] S. Wang *et al.*, "Linformer: Self-Attention with Linear Complexity," *arXiv:2006.04768*, 2020.

[67] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *arXiv:2103.14030*, 2021.

[68] M. Atad *et al.*, "CheXplaining in Style: Counterfactual Explanations for Chest X-rays using StyleGAN," *arXiv:2207.07553*, 2022.

[69] M. D. Zeiler *et al.*, "Visualizing and understanding convolutional networks," in *ECCV, Lect Notes Comput Sci*, vol. 8689. Springer, 2014, pp. 818–833.

[70] J. Castro *et al.*, "Polynomial calculation of the shapley value based on sampling," *Comput Oper Res*, vol. 36, no. 5, pp. 1726–1730, 2009.

[71] R. L. Buckner *et al.*, "The organization of the human cerebellum estimated by intrinsic functional connectivity," *J Neurophys*, vol. 106, no. 5, pp. 2322–2345, 2011.

[72] Y. Kamitani *et al.*, "Decoding the visual and subjective contents of the human brain," *Nat Neurosci*, vol. 8, no. 5, pp. 679–685, 2005.

[73] T. Naselaris *et al.*, "Encoding and decoding in fMRI," *NeuroImage*, vol. 56, no. 2, pp. 400–410, 2011.

[74] S. J. Ritchie *et al.*, "Sex differences in the adult human brain: evidence from 5216 UK biobank participants," *Cereb Cortex*, vol. 28, no. 8, pp. 2959–2975, 2018.

[75] X. Zhang *et al.*, "Gender differences are encoded differently in the structure and function of the human brain revealed by multimodal mri," *Front Hum Neurosci*, vol. 14, p. 244, 2020.

[76] E. McGlade *et al.*, "Sex differences in orbitofrontal connectivity in male and female veterans with tbi," *Brain Imaging Behav*, vol. 9, no. 3, pp. 535–549, 2015.

[77] S. Weis *et al.*, "Sex Classification by Resting State Brain Connectivity," *Cereb Cortex*, vol. 30, no. 2, pp. 824–835, 2019.

[78] D. Tomasi *et al.*, "Laterality patterns of brain functional connectivity: gender effects," *Cereb Cortex*, vol. 22, no. 6, pp. 1455–1462, 2012.

[79] B. Du *et al.*, "fMRI Brain Decoding and Its Applications in Brain-Computer Interface: A Survey." *Brain Sci*, vol. 12, no. 2, p. 228, 2022.

[80] Y. Fan *et al.*, "Multivariate examination of brain abnormality using both structural and functional MRI," *NeuroImage*, vol. 36, no. 4, pp. 1189–1199, 2007.

[81] L. Zhang *et al.*, "Deep fusion of brain structure-function in mild cognitive impairment," *Med Image Anal*, vol. 72, p. 102082, 2021.

[82] I. Kiremitci *et al.*, "Attentional Modulation of Hierarchical Speech Representations in a Multitalker Environment," *Cereb Cortex*, vol. 31, no. 11, pp. 4986–5005, 2021.

[83] D. Wang *et al.*, "CTformer: convolution-free Token2Token dilated vision transformer for low-dose CT denoising," *Phys Med Biol*, vol. 68, no. 6, p. 065012, 2023.

[84] O. Dalmaz *et al.*, "ResViT: Residual vision transformers for multimodal medical image synthesis," *IEEE Trans Med Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

[85] Z. Kong *et al.*, "DiffWave: A versatile diffusion model for audio synthesis," *arXiv:2009.09761*, 2021.

[86] M. Özbey *et al.*, "Unsupervised medical image translation with adversarial diffusion models," *IEEE Trans Med Imaging*, vol. 42, no. 12, pp. 3524–3539, 2023.

[87] Z.-X. Cui *et al.*, "Meta-learning enabled score-based generative model for 1.5t-like image reconstruction from 0.5t mri," *arXiv:2305.02509*, 2023.

[88] D. Garibi *et al.*, "Renoise: Real image inversion through iterative noising," *arXiv:2403.14602*, 2024.

[89] W. Xia *et al.*, "Low-dose CT using denoising diffusion probabilistic model for 20× speedup," *arXiv:2209.15136*, 2022.

[90] G. H. Ngo *et al.*, "A transformer-Based neural language model that synthesizes brain activation maps from free-form text queries," *Med Image Anal*, vol. 81, p. 102540, 2022.

[91] A. Craik *et al.*, "Deep learning for electroencephalogram (EEG) classification tasks: a review." *J Neural Eng*, vol. 16, no. 3, p. 031001, 2019.

[92] J. Aoe *et al.*, "Automatic diagnosis of neurological diseases using MEG signals with a deep neural network," *Sci Rep*, vol. 9, no. 1, p. 5057, 2019.

[93] C. Eastmond *et al.*, "Deep learning in fNIRS: a review." *Neurophotonics*, vol. 9, no. 4, p. 041411, 2022.

[94] A. P. Buccino *et al.*, "Combining biophysical modeling and deep learning for multielectrode array neuron localization and classification." *J Neurophysiol*, vol. 120, no. 3, pp. 1212–1232, 2018.

[95] W. Yan *et al.*, "Deep Learning in Neuroimaging: Promises and challenges," *IEEE Sig Process Mag*, vol. 39, no. 2, pp. 87–98, 2022.