Autoregressive Image Diffusion: Generation of Image Sequence and Application in MRI

Guanxiong Luo
University Medical Center Göttingen
guanxiong.luo@med.uni-goettingen.de

Shoujin Huang
Shenzhen Technology University

Martin Uecker
Graz University of Technology
uecker@tugraz.at

Abstract

Magnetic resonance imaging (MRI) is a widely used non-invasive imaging modality. However, a persistent challenge lies in balancing image quality with imaging speed. This trade-off is primarily constrained by k-space measurements, which traverse specific trajectories in the spatial Fourier domain (k-space). These measurements are often undersampled to shorten acquisition times, resulting in image artifacts and compromised quality. Generative models learn image distributions and can be used to reconstruct high-quality images from undersampled k-space data. In this work, we present the autoregressive image diffusion (AID) model for image sequences and use it to sample the posterior for accelerated MRI reconstruction. The algorithm incorporates both undersampled k-space and pre-existing information. Models trained with fastMRI dataset are evaluated comprehensively. The results show that the AID model can robustly generate sequentially coherent image sequences. In MRI applications, the AID can outperform the standard diffusion model and reduce hallucinations, due to the learned inter-image dependencies. The project code is available at `https://github.com/mrirecon/aid`.

1 Introduction

Magnetic resonance imaging (MRI) is a non-invasive imaging modality widely used in clinical practice to visualize soft tissue. Despite its utility, a persistent challenge in MRI is the trade-off between image quality and imaging speed. The trade-off is influenced by the k-space (spatial Fourier domain) measurements, which traverse spatial frequency data points along given sampling trajectories. To reduce acquisition time, the k-space measurements are often undersampled, resulting in image artifacts and reduced image quality.

In recent years, deep learning-based methods have emerged to improve image reconstruction in MRI. These methods are formulated as an inverse problem building upon compressed sensing techniques [1, 2] and benefit from the learned prior information instead of hand-crafted priors [3–5]. Another successful approach involves learning an image prior parameterized by a generative neural network [6, 7], which is then used as the learned and decoupled regularization on the image. Generative priors offer flexibility in handling changes in the forward model and perform well in reconstructing high-quality images from undersampled data.

Diffusion models [8–10], a class of generative models, have gained attention in recent years and are making an impact in many fields, including MRI reconstruction [11, 12]. These models learn to reverse a diffusion process that transforms random noise into structured images, producing high-

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-4100

quality, detailed images. Various approaches, including denoising diffusion probabilistic models (DDPMs) [10], denoising score matching [9], and continuous formulations based on stochastic differential equations (SDEs) [13], have been proposed for deriving diffusion models.

Recent studies demonstrate the effectiveness of diffusion models in accelerated MRI and their flexibility in handling various sampling patterns [11, 14–17]. For example, training score-based generative models using Langevin dynamics yields competitive reconstruction results for both in-distribution and out-of-distribution data [11]. Additionally, score-based diffusion models trained solely on magnitude images can reconstruct complex-valued data [15]. Comprehensive approaches using data-driven Markov chains facilitate efficient MRI reconstruction across variable sampling schemes and enable the generation of uncertainty maps [16].

Autoregressive models are statistical models that predict the current value of a variable based on its past values, capturing temporal dependencies and patterns within the data. They are widely used in various fields such as time series analysis, signal processing, and sequence modeling. In natural language processing, autoregressive models like generative pre-trained transform (GPT) [18, 19] predict each token in a sequence based on previously generated tokens, enabling the generation of coherent and contextually relevant text. Similarly, in image modeling, autoregressive models like PixelCNN [20] and ImageGPT [21] generate images by predicting each pixel value based on previously generated pixel values, often in a left-to-right, top-to-bottom order. Instead of directly modeling pixels, which can be computationally expensive for high-resolution images, the study [22] proposes to first compress the image into a smaller representation using vector quantized variational autoencoder (VQVAE). This VQVAE learns a codebook of visually meaningful image components. Then, a transformer is applied to model the autoregressive relationship between these components, effectively capturing the global structure of the image. By predicting each image component based on previous ones, the model generates high-resolution images in a sequential manner, maintaining consistency and coherence across the entire image.

The clinical practice of MRI often involves acquiring volumetric image sequences to monitor disease progression and treatment response; modeling and generating these image sequences is challenging. Autoregressive models can be employed to model the joint distribution of image sequences and extract the dependencies between images. The diffusion process is effective in modelling images by treating each image independently. Therefore, we aim to combine these two models and propose autoregressive image diffusion (AID) model to generate sequences of images.

The contributions of this work are the following aspects. We present how to derive the autoregressive image diffusion training loss starting from a common diffusion loss and how to optimize loss in parallel for efficient training. We present the algorithm to sample the posterior for accelerated MRI reconstruction when using AID to facilitate the incorporation of pre-existing information. We performed experiments to evaluate its ability in generating images when different the amount of initial information is given and to validate its effectiveness in MRI reconstruction. The results show that the AID model can stably generate highly coherent image sequences even without any pre-existing information. When used as a generative prior in MRI reconstruction, the AID outperforms the standard diffusion model and reduces the hallucinations in the reconstructed images, benefiting from the learned prior knowledge about the relationship between images and pre-existing information.

2 Methods

2.1 Autoregressive image diffusion

Given a dataset X consisting of multiple sequences of images, each sequence represented as $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$, our goal is to model the joint distribution of these images. This joint distribution is autoregressively factorized into the product of conditional probabilities:

$$p(\mathbf{x}) = q(x_1|x_0) \prod_{t=2}^{N} q(x_n|x_{<n}), \tag{1}$$

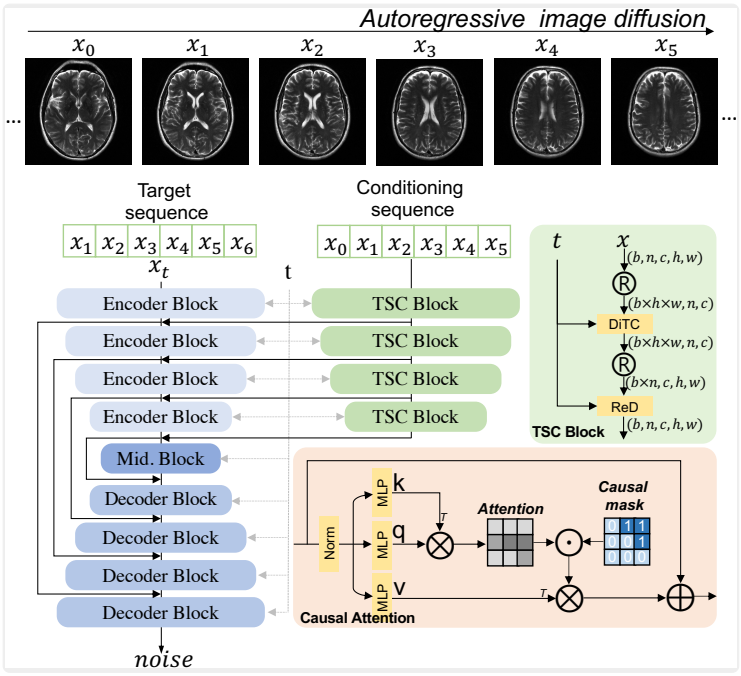


Figure 1: The interaction between the images in conditioning sequence occurs in the DiTBlock, which has a causal attention module to ensure x_n is conditioned on previous images $x_{<n}$. During training, the net predicts the noise for each noisy image that is sampled from the target sequence given the conditioning sequence in parallel. During generation, the net iteratively refines the noisy input to produce a clean image, which is then appended to the conditioning sequence.

where $x_{<n} = \{x_1, x_2, \ldots, x_{n-1}\}$ and the image x_0 is known. The model parameterized by θ is trained by minimizing the negative log-likelihood of the data:

$$\mathcal{L}_{AID} = \mathbb{E}_X \left[-\log p_\theta(\mathbf{x}) \right] = \mathbb{E}_X \left[-\log p_\theta(x_1|x_0) - \sum_{t=2}^{N} \log p_\theta(x_n|x_{<n}) \right]. \tag{2}$$

Sohl-Dickstein et al. (2015) and Ho et al. (2020) introduced the denoising diffusion probabilistic model (DDPM). This model gradually introduces fixed Gaussian noise to an observed data point x^0 using known scales β_t, generating a series of progressively noisier values $x^1, x^2, \ldots, x^T$. The final noisy output x^T follows a Gaussian distribution with zero and identity covariance matrix I, containing no information about the original data point. The series of positive noise scales $\beta_1, \ldots, \beta_T$ must be increasing, ensuring that the first noisy output x^1 closely resembles the original data x^0, while the final value x^T represents pure noise. We apply this process to the conditional probability $q(x_n|x_{<n})$ in Equation (2) by adding the noise to the image independent of the position in the sequence, i.e., x_n^t and $x_{<n}^0$ are conditionally independent given x_n^{t-1}. Then the transition from x_n^{t-1} to x_n^t is defined as:

$$q(x_n^t|x_n^{t-1}, x_{<n}^0) = q(x_n^t|x_n^{t-1}) = \mathcal{N}(x_n^t; \sqrt{1-\beta_t}x_n^{t-1}, \beta_t \mathbf{I}) \tag{3}$$

Here, x_n^t represents the image x_n at time t, x_n^{t-1} is the image at the previous time step, and $x_{<n}^0$ denotes all images preceding x_n at the initial time step. The parameter β_t controls the drift and diffusion of this process. The objective is to learn to reverse this process. The reverse process is defined as:

$$p_\theta(x_n^{t-1}|x_n^t, x_{<n}^0) = \mathcal{N}(x_n^{t-1}; \mu_\theta(x_n^t, x_{<n}^0, t), \Sigma_\theta(x_n^t, x_{<n}^0, t)), \tag{4}$$

where μ_θ and Σ_θ are parameterized by a neural network θ, taking x_n^t, $x_{<n}^0$, and t as inputs. Using the variational lower bound, the reverse process can be learned by minimizing the negative log-likelihood of the data:

$$\mathbb{E}[-\log p_\theta(x_n|x_{<n}^0)] \leq \mathbb{E}\left[-\log p(x_n^T) - \sum_{t\geq 1} \log \frac{p_\theta(x_n^{t-1}|x_n^t, x_{<n}^0)}{q(x_n^t|x_n^{t-1}, x_{<n}^0)} \right] := L_{D_n}, \tag{5}$$

https://doi.org/10.52202/079017-4100

Given the initial image x_n^0 and that x_n^t and $x_{<n}^0$ are conditionally independent given x_n^0, x_n^t at an arbitrary time step t is sampled from a Gaussian distribution:

$$q(x_n^t|x_n^0, x_{<n}^0) = \mathcal{N}(x_n^t; \sqrt{\bar{\alpha}_t}x_n^0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{6}$$

using $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The posterior distribution x_n^{t-1} given x_n^0 and x_n^t is then calculated as:

$$q(x_n^{t-1}|x_n^t, x_n^0, x_{<n}^0) = \mathcal{N}(x_n^{t-1}; \tilde{\mu}_t(x_n^t, x_n^0), \tilde{\beta}_t\mathbf{I}), \tag{7}$$

where $\tilde{\mu}_t(x_n^t, x_n^0) := \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_n^0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_n^t$ and $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$.

The training objective Equation (5) is further written as minimizing the Kullback-Leibler (KL) divergence between the forward and reverse processes in Equation (4) and Equation (7), as proposed by Sohl-Dickstein et al. (2015). (See Appendix A for details.)

In practice, the approach proposed by Ho et al. (2020) involves reparameterizing μ_θ and predicting the noise ϵ for x_n^t. The expression for x_n^t is given by $x_n^t(x_n^0, \epsilon) = \sqrt{\bar{\alpha}_t}x_n^0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, with $\Sigma_\theta(x_n^t, x_{<n}^0, t) = \beta_t$ fixed. We realized this with a neural network $\epsilon_\theta(x_n^t, t, x_{<n}^0)$ shown in Figure 1, which predicts the noise for x_n^t at each time step given $x_{<n}^0$. In the end, the objective function in Equation (2) for training autoregressive image diffusion is written as

$$\mathcal{L}_{AID} \geq \sum_{n=1}^{N} L_{D_n} = \sum_{n=1}^{N} \mathbb{E}_{t,\epsilon|x_n^0,x_{<n}^0}\left[\left\|\epsilon_\theta(\sqrt{\bar{\alpha}_t}x_n^0 + \sqrt{1-\bar{\alpha}_t}\epsilon, x_{<n}^0, t) - \epsilon\right\|_2^2\right], \tag{8}$$

where the expectation is taken over the noise $\epsilon \sim \mathcal{N}(0, I)$ and the time step $t \sim \mathcal{U}(1, ..., T)$. To generate an image sequence, we begin with the noise x_1^T and update it iteratively using Equation (4) with the given x_0^0, following the sequence $(x_1^T \to x^{T-1} \to \ldots \to x_1^0)$. This process yields a clean sample x_1^0. Subsequently, we can sample x_2^0 in the same manner using the generated images $x_{<2}^0$, and continue this process iteratively to generate the entire sequence of images.

2.2 Architecture

To optimize the objective function in Equation (8) efficiently, ordered images are loaded as sequences of a certain length $N + 1$ during the training phase. We take the first N images $\mathbf{x}_{con} = \{x_0, x_1, ..., x_{N-1}\}$ as the conditioning sequence and the last N images $\mathbf{x}_{target} = \{x_1, ..., x_N\}$ as the target sequence, as shown in Figure 1. We adopt an architecture built on an Unet [23] with capabilities of temporal-spatial conditioning (TSC), designed to process the conditioning sequence and predict the noise for the target sequence. The term "temporal" refers to conditioning in previous frames along the N dimensions, while the "spatial" refers to the conditioning in the previous frame among the $H \times W$ dimensions. Additionally, the TSC block is conditioned on the time steps t of the diffusion process.

The only interaction between images in the conditioning sequence occurs during the attention operation. To maintain proper conditioning with autoregressive property, we implemented a standard upper triangular mask on the $n \times n$ matrix of attention logits. This causal attention module is used in DiTBlock [18, 24]. The modified DiTBlock is followed by a ResNet block [25], which is a standard building block in the Unet architecture. The features output by the TSC block are then passed to the corresponding encoder block in the Unet, which process the target sequence. The change in tensor dimensions inside TSC Block is handled by the einops library[1] and illustrated in Figure 1.

During training, the net predicts the noise in parallel for each noisy image that is sampled from the target sequence, given the conditioning sequence. During generation of sequence, the net iteratively refines the noisy input to produce a clean image, which is then appended to the conditioning sequence.

2.3 Application in MRI inverse problem

Image reconstruction is formulated as a Bayesian problem where the posterior of image $p(x|y)$ is expressed as

$$p(x|y) = \frac{p(y|x) \cdot p(x)}{p(y)}\,. \tag{9}$$

[1] https://github.com/arogozhnikov/einops

Here, y represents the measured k-space data, x denotes the image, and $p(x)$ is a generative prior. The minimum mean square error (MMSE) estimator for the posterior minimizes the mean square error, given by:

$$x_{\text{MMSE}} = \arg\min_{\tilde{x}} \int \|\tilde{x} - x\|^2 p(x|y)dx = \mathbb{E}[x|y] \, . \tag{10}$$

2.4 Likelihood function for k-space

The image $x \in \mathbb{C}^{n\times n}$ is represented as a complex matrix , where $n \times n$ is the image size, and $y \in \mathbb{C}^{m\times m_C}$ is a vector of complex-valued k-space samples from m_C receive coils. Assuming circularly-symmetric normal noise η with zero mean and covariance matrix $\sigma_\eta^2\mathbf{I}$, the likelihood $p(y|x)$ of observing y given x is formulated as a complex normal distribution:

$$\begin{aligned} p(y|x) &= \mathcal{CN}(y; \mathcal{A}x, \sigma_\eta^2\mathbf{I}) \\ &= (\sigma_\eta^2\pi)^{-N_p} e^{\text{-}\|\sigma_\eta^{-1}\cdot(y-\mathcal{A}x)\|_2^2} \, , \end{aligned} \tag{11}$$

where $\mathbf{I}$ is the identity matrix, σ_η is the standard deviation of the noise, $\mathcal{A}x$ represents the mean, and N_p is the length of the k-space data vector. The operator $\mathcal{A} : \mathbb{C}^{n\times n} \to \mathbb{C}^{m\times m_C}$ maps the image x to k-space and is composed of the coil sensitivity maps $\mathcal{S}$, the two-dimensional Fourier transform $\mathcal{F}$, and the k-space sampling mask $\mathcal{P}$, defined as $\mathcal{A} = \mathcal{PFS}$. For more details and visual understanding on the forward operator, please refer to Appendix C.

2.5 Sampling the posterior

Given a sequence of k-space $\mathbf{y} = \{y_1, \ldots, y_N\}$, each posterior in $\{p_\theta(x_n|y_n, x^0_{<n})|1 < n < N\}$ is expressed as

$$\begin{aligned} p_\theta(x_n|y_n, x^0_{<n}) &= \frac{p(y_n|x_n, x^0_{<n})p_\theta(x_n|x^0_{<n})}{p(y_n|x^0_{<n})} = \frac{p(y_n|x_n)p_\theta(x_n|x^0_{<n})}{p(y_n)} \\ &\propto p(y_n|x_n)p_\theta(x_n|x^0_{<n}) \, , \end{aligned} \tag{12}$$

when the acquisition of y_n is independent of the image $x^0_{<n}$, y_n and $x^0_{<n}$ are conditionally independent given x_n. Following the Reference [8], we have

$$p_\theta(x_n^{t-1}|x_n^t, y_n, x^0_{<n}) \propto p(y_n|x_n^t)p_\theta(x_n^{t-1}|x^t_{<n}, x^0_{<n}) \, . \tag{13}$$

The details for Equation (13) is in Appendix B. To sample the above posterior, the learned reverse process in Equation (4) is used, and the algorithm is constructed with two gradient updates using the log of the prior and k-space likelihood: the DDIM (Denoising Diffusion Implicit Model) reverse step proposed by Song et al. (2020), and a data fidelity step derived from the likelihood function Equation (11), which are described as follows:

$$\tilde{x}_n^{t-1} \leftarrow \sqrt{\alpha_{t-1}}\left(\frac{x_n^t - \sqrt{1-\alpha_t}\epsilon_\theta(x_n^t, x^0_{<n}, t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta(x_n^t, x^0_{<n}, t) \tag{14}$$

$$x_n^{t-1} \leftarrow \tilde{x}_n^{t-1} + \lambda \cdot \nabla_{x_n^{t-1}} \log p(y_n|\tilde{x}_n^{t-1}) \, . \tag{15}$$

where λ is the step size, and $\nabla_{x_n^{t-1}} \log p(y_n|x_n^{t-1})$ is the gradient of the log-likelihood of Equation (11). Then, the reconstruction of a sequence images from the undersampled k-space data is achieved by sequentially sampling the posterior in $\{p(x_n|y_n, x^0_{<n})|1 < n < N\}$ using autoregressive diffusion model as prior. The algorithm is summarized in Algorithm 1.

3 Experiments and Results

3.1 Model training

Two autoregressive diffusion models were trained on separate datasets: one in image space and the other in latent space. The image space model was trained on brain images that are from the fastMRI training dataset, which includes T1-weighted (some with post-contrast), T2-weighted, and FLAIR images [27]. These complex images were reconstructed from fully sampled multi-channel k-space

 https://doi.org/10.52202/079017-4100

Algorithm 1 Sample the posterior in $\{p(x_n|y_n, x^0_{<n})|1 < n < N\}$ using autoregressive diffusion model as prior.

Initial image sequence: $x^0_{<n} = x_0$; Time steps: T; Step size: λ; Iterations for data fidelity step: K; Number of samples: S;
for y_n in $\mathbf{y} = \{y_1, y_2, ..., y_N\}$ **do**
    Initialize x^T_n with Gaussian noise.
    Construct the forward operator $\mathcal{A}$ with sampling pattern $\mathcal{P}$ and coil sensitivities $\mathcal{S}$.
    for t in $\{T-1, \ldots, 0\}$ **do**
        Run the DDIM reverse step in Equation (14) to get x^{t-1}_n given x^t_n and $x^0_{<n}$.
        Run the data fidelity step in Equation (15) to update x^{t-1}_n for K step.
        Add Gaussian noise scaled by $\sqrt{1-\alpha_{t-1}}$ to x^{t-1}_n.
    end for
    Update $x^0_{<n} \leftarrow \{x^0_n, \ldots, x^0_0\}$.
end for

volumes, with coil sensitivity maps computed using the BART toolbox [28]. The images were then normalized to a maximum magnitude of 1, and the real and imaginary parts were treated as separate channels when input into the neural network. The number of images in each volume ranged from 12 to 16. Images were loaded without reordering and resized to 320×320 pixels if they were not already that size.

The latent space model is trained with the cardiac dataset that contains cine images reconstructed by the SSA-FARY method [29]. Firstly, a VQVAE was trained on the cine images that were preprocessed similarly to images in fastMRI dataset. The cine images have a size of 256×256 pixels. Then, it generates latent space for the training AID. (See the details for configuration of VQVAE in Appendix J). All the training was performed on 4 NVIDIA A100 GPUs with 80GB memory. The models were trained using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 1 for image space model and 4 for latent space model. Two models were trained for 440,000 iterations. It took around 2 hours to train brain model for 10k steps and 1.2 hours for cardiac model. The length of conditioning sequence N for brain and cardiac models are 10 and 42. The network as illustrated in Figure 1 was implemented based on OpenAI's guided diffusion codebase[2]. We also trained a standard diffusion model, Guide, on the brain dataset for comparison. The Guide model was trained using the same hyperparameters as the AID model, except the batch size is 10. The Guide model uses the same Unet blocks as AID.

3.2 Generating sequence of images

To test different aspects of the autoregressive diffusion models, we generate the sequence of images using the following two approaches.

Retrospective sampling: This method generates a new sequence of images $\{\tilde{x}_1, \ldots, \tilde{x}_N\}$ based on the given sequence $\{x_0, \ldots, x_{N-1}\}$. $\tilde{x}_n$ is sampled from Equation (4) given $\{x_0, \ldots, x_{n-1}\}$.

Prospective sampling: A fixed-length sliding window is initialized with the given sequence $x_{<N} = \{x_0, \ldots, x_{N-1}\}$. x_N is generated from Equation (4) with the current window as conditioning. Subsequently, the window is updated by appending the newly generated x_N and removing the earliest image x_0. This autoregressive sampling process is repeated until the stop condition is met. We refer to this process as a warm start. In a cold start, the window is initialized with zeros, and each element x_n in it is updated with newly generated images from the beginning to the end, after which the generation is warmed up.

In the retrospective sampling, the model generates a sequence of images that are sequentially coherent and visually similar to the conditioning sequence, as shown in Figure 2 (a). The prospective sampling generates a sequence of images that extends the initial images in the sliding window and constitutes multiple volumes, as shown in Figure 2 (b). As for a cold start, Figure 3 demonstrates the model's ability to generate a sequence of images using black background as initial status. This shows the model's generative capabilities from a minimal initial condition, thereby proving its robustness and flexibility. Due to the limit of space, the samples with similar quality from the model trained on

[2]https://github.com/openai/guided-diffusion

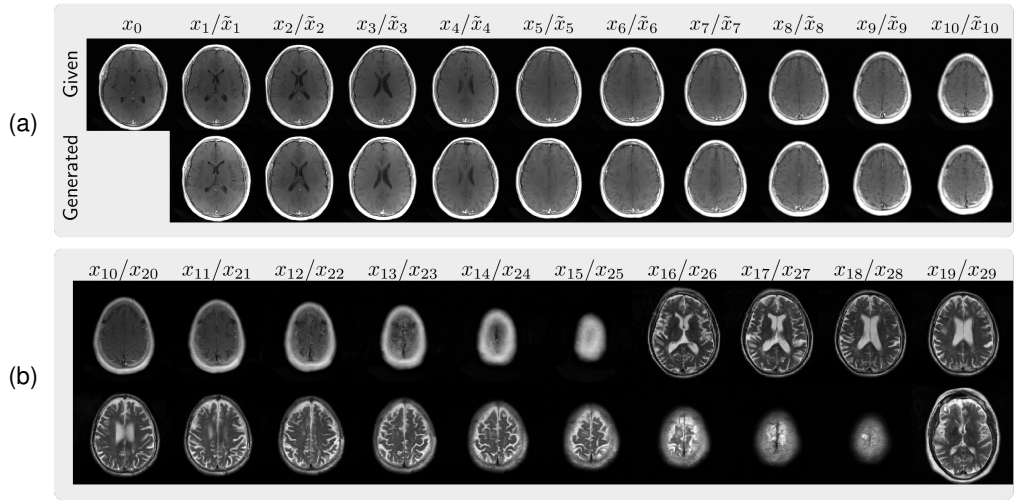


Figure 2: (a): A sequence of images from dataset is shown in the first row and is used as conditioning to generate retrospective samples that are shown in the second row. (b): With the given sequence in (a) as a warm start, prospective samples extending it are shown.

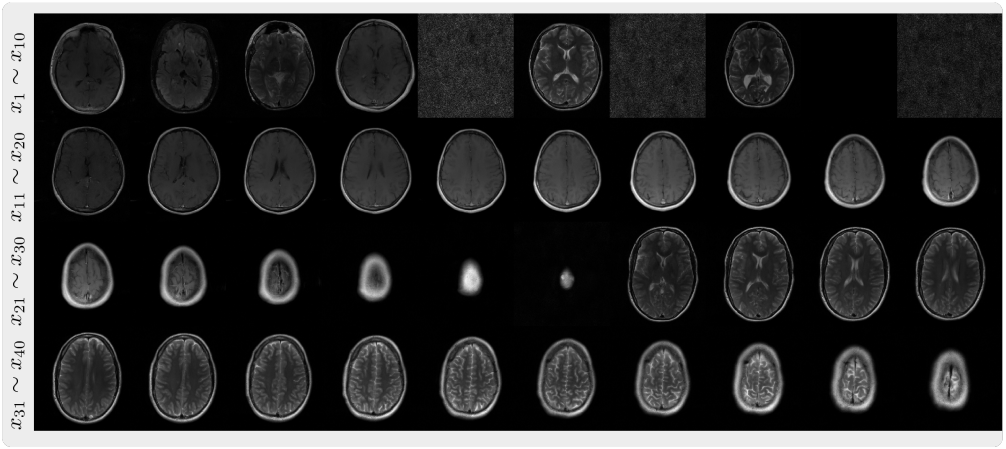


Figure 3: Prospective samples with cold start. The initial images generated in the cold start are not sequentially coherent, but as the sampling process continues, the model progressively generates more sequentially coherent and realistic images.

the cardiac dataset are shown in Appendix D. We also implemented a boosted sampling technique which use previous slice with added noise as the initial image for the current slice. This requires less iterations to generate the sequence of images. Further details can be found in our codebase.

3.3 MRI reconstruction

The MMSE estimator in Equation (10) cannot be computed in a closed form, and numerical approximations are typically required. Once the samples from the posterior is obtained with Algorithm 1, a consistent estimate of $x_{n\mathrm{MMSE}}$ can be computed by averaging those samples, i.e. the empirical mean of samples converges in probability to $x_{n\mathrm{MMSE}}$ due to weak law of large numbers. The variance of those samples provides a solution to the error assessment in the reconstruction assuming the trained model is trusted. To highlight the regions with large uncertainty, we compute the pseudo-confidence intervals based on the assumption that each pixel's intensity is normally distributed. This involves determining the standard error from the variance, then multiplying it by the t-score corresponding to a 95% confidence level.

 https://doi.org/10.52202/079017-4100

Unfolding of aliased single-coil image: To investigate how the trained model, AID, reduces the folding artifacts in the reconstruction, we designed the single coil unfolding experiment. The single-channel k-space is simulated out of multichannel k-space data. The odd lines in k-space are retained, y. Ten samples were drawn from the posterior $p(x_1|y, x_0)$ using Algorithm 1 with parameters: $T = 1000, \lambda = 1, K = 5$. The experiment was repeated using a standard diffusion model, Guide. The results are shown in Figure 4. The AID model significantly reduces the errors in the region of folding artifacts compared to the Guide model. The mean over samples, x_{MMSE}, is highlighted with a confidence interval computed from the variance of samples. The highlighted mean image shows the reconstruction by AID is more trustworthy in the folding region. In general, the highlighted region lies in the folding region, where large errors remains, as we expected.

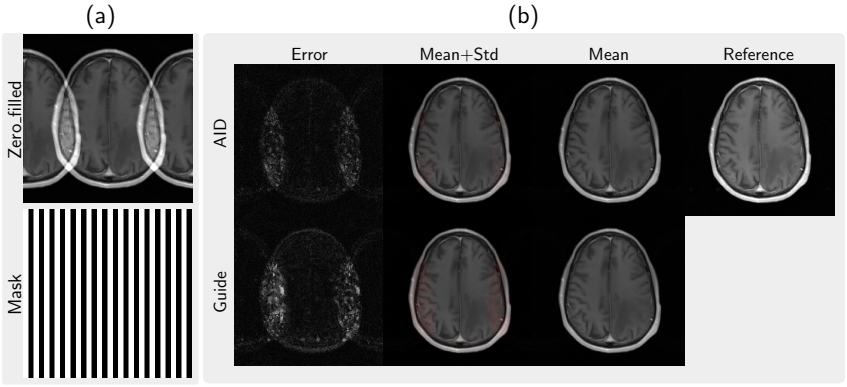


Figure 4: (a): The folded single-coil image caused by two-times undersampling mask. (b): The comparison of unfolding ability by the autoregressive and the standard diffusion model, i.e., AID (top) and Guide (bottom). Reference image is reconstructed from k-space data without undersampling. The error is the difference between the mean, x_{MMSE}, and the reference image. The "Mean+std" is the mean highlighted with confidence interval, which indicates the reconstruction by AID is more trustworthy in the region of folding artifacts.

Reconstruction from undersampled data: To further investigate the model's performance in reconstruction, we conducted experiments on 20 volumes from the fastMRI validation dataset where k-space data was retrospectively undersampled using various sampling masks. We created four types of sampling masks: random with autocalibration signal (ACS), random without ACS, equispaced with ACS, and equispaced without ACS. The undersampling factor is 12. Setting parameters: $T = 1000, \lambda = 1, K = 4$ for Algorithm 1, the images were reconstructed from the undersampled k-space data using the AID and Guide as prior, respectively. Another method proposed in Ref. [11] is used as the baseline (CSGM), which uses a scored-based model (NCSNv2) from Ref. [30] trained on the fastMRI dataset. All the reconstruction tasks are performed by sampling the posterior. The likelihood $p(y|x)$ is determined by forward model and the image prior is determined by the trained models, such as NCSNv2, Guide, and AID. This means that when the sampling method remains consistent, the performance of the reconstruction task is determined by the quality of the image prior. Our algorithm treats $p(y|x)$ in the same manner, and the key difference is the image prior.

We used peak-signal-noise-ratio (PSNR in dB) and normalized root-mean-square error (NRMSE) to evaluate the reconstruction quality against the reference image that is reconstructed from full k-space. The comparison of metrics across experimental conditions is illustrated in Figure 5. The proposed AID model outperforms the Guide and NCSNv2 in terms of PSNR and NRMSE especially in the absence of ACS, demonstrating its superior performance in image reconstruction from undersampled k-space data. The results are consistent across different undersampling factors and sampling masks, indicating the model's robustness and flexibility in handling various types of undersampled k-space data.

For the visual impression of the improvement by the AID model in reconstruction, we show the reconstructed images in Figure 6 and more of them in Appendix E. The images reconstructed using AID are more visually similar to the reference images than using Guide, even which also provides aliased-free images. Furthermore, it is worth noting that more visually notable hallucinations were introduced by the Guide model than the AID model, which means AID is more trustworthy.

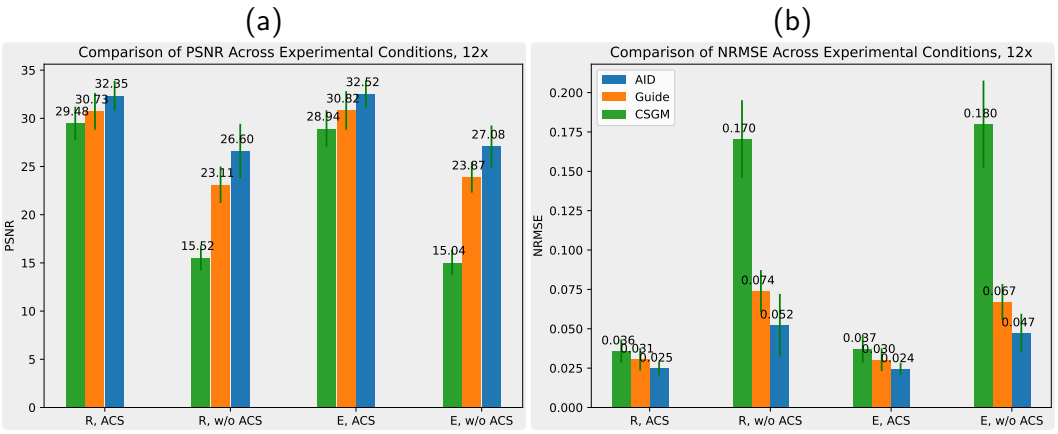


Figure 5: E: equispaced, R: random. (a): PSNR and (b): NRMSE of the images reconstructed from the twelve-times undersampled k-space data using the autoregressive diffusion model (AID), the standard diffusion model (Guide), and the baseline method CSGM. PSNR higher is better, and NRMSE lower is better.

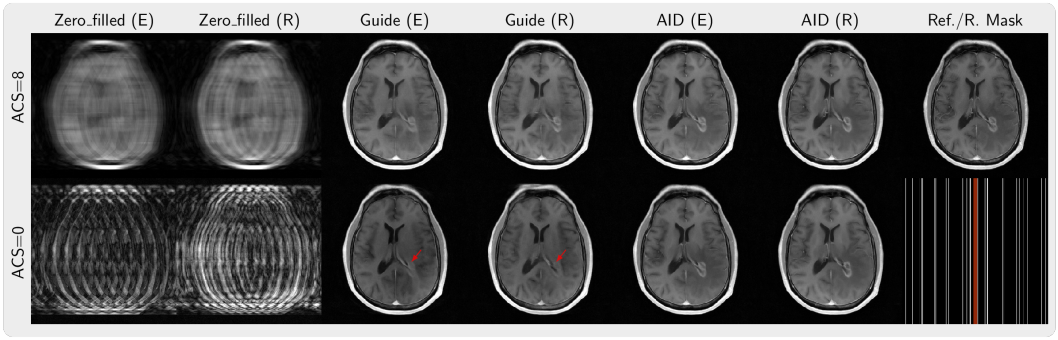


Figure 6: E: equispaced, R: random. The last column shows the reference and the random sampling mask in k-space. The red lines are autocalibration signal (ACS) and equispaced mask is not shown. Zero-filled images are computed by inverse Fourier transform of the zero-filled k-space data. The hallucinations are pointed with red arrows.

3.4 Other models for image sequence generation

To further evaluate the model's ability to generate image sequences, we further trained two AID models on two datasets: one using brain images from autism studies called ABIDE [31, 32], and the other using images from Unmanned Aerial Vehicle (UAV) view dataset [33]. We reported the computation and model complexity in Appendix F for all AID models trained in this work. We also implemented a two-stage training to improve the efficiency for training models on ABIDE and presented correspondingly generated samples in Appendix G. We demonstrated the natural image sequence generation in Appendix H and showed the sample consistency along the temporal axis in Appendix I.

4 Discussion

In this work, we propose an autoregressive image diffusion model for generating image sequences, with specific applications to accelerated MRI reconstruction. We conducted comprehensive evaluation of its performance as an image prior in reconstruction algorithms, comparing it to a standard diffusion model. Due to the learned prior information on inter-image dependencies, the proposed model outperforms the standard diffusion model across various scenarios. Our model is particularly well-

https://doi.org/10.52202/079017-4100

suited for medical applications where image sequences are often acquired (e.g., in volumetric format) from patients in clinical practice. For instance, when different contrast images are acquired during an examination session [34], our model is designed to capture the relationships between these images. This enables more accurate and coherent reconstructions from undersampled k-space data using the proposed Algorithm 1. Additionally, other medical imaging tasks like dynamic MRI, multi-contrast, super-resolution, and denoising could benefit from our model's ability by leveraging inter-image dependencies [35]. Furthermore, the proposed algorithm holds great promise for facilitating the incorporation pre-existing information from other imaging modalities into MRI image reconstruction. This opens up a wide range of potential medical applications, with the potential to improve patient care and reduce healthcare costs by enabling faster and more accurate image acquisition and diagnosis.

Privacy Issue: As this model has the capability to generate coherent image sequences, it is crucial to consider the privacy implications associated with its use, particularly in clinical settings. The generation of such images may inadvertently expose sensitive patient information, including identifiable features such as facial characteristics. Safeguarding patient privacy must be a top priority when deploying it. We recommend that the model be used in a controlled environment where access to the generated images is restricted to authorized personnel only. Additionally, it is essential to ensure that the model is trained on anonymized data and that the generated images are not stored or shared without proper consent.

Limitation and future work: We did not evaluate the model on a common image dataset such as ImageNet or Cifar-10, nor did we compute metrics such as FID and Inception Score, which could be a limitation of our work. We plan to address these limitations in future work by running the model on a large dataset and comparing it with other state-of-the-art models. Additionally, given the model's suitability for modeling image sequences, it is worth exploring its potential for optimizing MRI k-space acquisition strategies, as the acquisition process constitutes a sequence of operations.

5 Conclusion

The proposed autoregressive image diffusion model offers an approach to generating image sequences, with significant potential as a trustworthy prior in accelerated MRI reconstruction. In various experiments, it outperforms the standard diffusion model in terms of both image quality and robustness by taking the advantage of the prior information on inter-image dependencies.

Acknowledgements

This work was supported by DZHK (German Centre for Cardiovascular Research) funding code: 81Z0300115. We acknowledge funding by the "Niedersächsisches Vorab" funding line of the Volkswagen Foundation. This work was supported by the Federal Ministry of Education and Research (BMBF), Germany under the AI service center KISSKI (grant no. 01IS22093A).

References

[1] M. Lustig, D. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.*, 58(6):1182–1195, 2007.

[2] K. T. Block, M. Uecker, and J. Frahm. Undersampled radial MRI with multiple coils. Iterative image reconstruction using a total variation constraint. *Magn. Reson. Med.*, 57(6):1086–1098, 2007. ISSN 1522-2594. doi: 10.1002/mrm.21236.

[3] Yan Yang, Jian Sun, Huibin Li, and Zongben Xu. Deep ADMM-Net for Compressive Sensing MRI. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[4] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P. Recht, Daniel K. Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magn. Reson. Med.*, 79(6):3055–3071, 2017. ISSN 1522-2594. doi: 10.1002/mrm.26977.

[5] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas S Vasanawala, Greg Zaharchuk, Lei Xing, and John M Pauly. Deep generative adversarial neural networks for compressive sensing mRI. *IEEE transactions on medical imaging*, 38(1):167–179, 2018.

[6] Kerem C Tezcan, Christian F Baumgartner, Roger Luechinger, Klaas P Pruessmann, and Ender Konukoglu. MR image reconstruction using deep density priors. *IEEE transactions on medical imaging*, 38(7):1633–1642, 2019. doi: 10.1109/TMI.2018.2887072.

[7] Guanxiong Luo, Na Zhao, Wenhao Jiang, Edward S. Hui, and Peng Cao. MRI reconstruction using deep bayesian estimation. *Magn. Reson. Med.*, 84(4):2246–2261, apr 2020. doi: 10.1002/mrm.28274.

[8] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[9] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11895–11907, 2019.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[11] Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. *Advances in Neural Information Processing Systems*, 34:14938–14954, 2021.

[12] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[13] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[14] Alper Güngör, Salman UH Dar, Şaban Öztürk, Yilmaz Korkmaz, Hasan A Bedel, Gokberk Elmas, Muzaffer Ozbey, and Tolga Çukur. Adaptive diffusion priors for accelerated mri reconstruction. *Medical Image Analysis*, 88:102872, 2023.

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

[16] Guanxiong Luo, Moritz Blumenthal, Martin Heide, and Martin Uecker. Bayesian mri reconstruction with joint uncertainty estimation using diffusion models. *Magnetic Resonance in Medicine*, 90(1):295–311, 2023.

[17] Martin Zach, Florian Knoll, and Thomas Pock. Stable deep mri reconstruction using generative priors. *IEEE Transactions on Medical Imaging*, 2023.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[19] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[20] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29, 2016.

[21] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.

 https://doi.org/10.52202/079017-4100

[22] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[24] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[27] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020.

[28] Moritz Blumenthal, Martin Heide, Christian Holme, Martin Juschitz, Bernhard Rapp, Philip Schaten, Nick Scholand, Jon Tamir, Christian Tönnes, and Martin Uecker. mrirecon/bart: version 0.9.00, December 2023. URL `https://doi.org/10.5281/zenodo.10277939`.

[29] Sebastian Rosenzweig, Nick Scholand, H Christian M Holme, and Martin Uecker. Cardiac and respiratory self-gating in radial mri using an adapted singular spectrum analysis (ssa-fary). *IEEE transactions on medical imaging*, 39(10):3029–3041, 2020.

[30] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

[31] A. Di Martino, C-G Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R. A. Mueller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, Marc Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, and M. P. Milham. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, June 2014. ISSN 1359-4184. doi: 10.1038/mp.2013.78.

[32] Adriana Di Martino, David O'connor, Bosi Chen, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Joshua H Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, Laura M E Blanken, Susan y Bookheimer, B. Blair Braden, Lisa Byrge, F. Xavier Castellanos, Mirella Dapretto, Richard Delorme, Damien A Fair, Inna Fishman, Jacqueline Fitzgerald, Louise Gallagher, R. Joanne Jao Keehn, Daniel P Kennedy, Janet E Lainhart, Beatriz Luna, Stewart H Mostofsky, Ralph-Axel Müller, Mary Beth Nebel, Joel T Nigg, Kirsten O'hearn, Marjorie Solomon, Roberto Toro, Chandan J Vaidya, Nicole Wenderoth, Tonya White, R. Cameron Craddock, Catherine Lord, Bennett L. Leventhal, and Michael Milham. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, 4:170010, March 2017.

[33] İbrahim Delibaşoğlu. Pesmod: Small moving object detection benchmark dataset for moving cameras. In *2022 7th International Conference on Frontiers of Signal Processing (ICFSP)*, pages 23–29. IEEE, 2022.

[34] Brett Levac, Ajil Jalal, Kannan Ramchandran, and Jonathan I Tamir. Conditional score-based reconstructions for multi-contrast mri. *arXiv preprint arXiv:2303.14795*, 2023.

[35] Guangyuan Li, Chen Rao, Juncheng Mo, Zhanjie Zhang, Wei Xing, and Lei Zhao. Rethinking diffusion model for multi-contrast mri super-resolution. *arXiv preprint arXiv:2404.04785*, 2024.

A Loss function derivation

Below is a derivation of Equation (5), the reduced variance variational bound for diffusion models. This adapted from Sohl-Dickstein et al. (2015) and Ho et al. (2020). We include it here only for completeness. In the forward process, x_n^t and $x_{<n}^0$ are conditionally independent given x_n^{t-1}.

$$L = \mathbb{E}_q \left[-\log p(x_n^T | x_{<n}^0) - \sum_{t>1} \log \frac{p_\theta(x_n^{t-1}|x_n^t, x_{<n}^0)}{q(x_n^t|x_n^{t-1}, x_{<n}^0)} - \log \frac{p_\theta(x_n^0|x_n^1, x_{<n}^0)}{q(x_n^1|x_n^0, x_{<n}^0)} \right] \tag{16}$$

$$= \mathbb{E}_q \left[-\log p(x_n^T | x_{<n}^0) - \sum_{t>1} \log \frac{p_\theta(x_n^{t-1}|x_n^t, x_{<n}^0)}{q(x_n^{t-1}|x_n^t, x_n^0, x_{<n}^0)} \cdot \frac{q(x_n^{t-1}|x_n^0)}{q(x_n^t|x_n^0)} - \log \frac{p_\theta(x_n^0|x_n^1, x_{<n}^0)}{q(x_n^1|x_n^0, x_{<n}^0)} \right] \tag{17}$$

$$= \mathbb{E}_q \left[-\log \frac{p(x_n^T|x_{<n}^0)}{q(x_n^T|x_n^0, x_{<n}^0)} - \sum_{t>1} \log \frac{p_\theta(x_n^{t-1}|x_n^t, x_{<n}^0)}{q(x_n^{t-1}|x_n^t, x_n^0, x_{<n}^0)} - \log \frac{p_\theta(x_n^0|x_n^1, x_{<n}^0)}{q(x_n^1|x_n^0, x_{<n}^0)} \right] \tag{18}$$

$$= \mathbb{E}_q \Bigg[D_{\mathrm{KL}}(q(x_n^T|x_n^0, x_{<n}^0) \parallel p(x_n^T|x_{<n}^0)) + \sum_{t>1} D_{\mathrm{KL}}(q(x_n^{t-1}|x_n^t, x_n^0) \parallel p_\theta(x_n^{t-1}|x_n^t, x_{<n}^0)) \tag{19}$$

$$- \log p_\theta(x_n^0|x_n^1, x_{<n}^0) \Bigg] \tag{20}$$

B Posterior derivation

When samples drawn from the posterior started from the standard Gaussian noise, with Equation (12) we have

$$p(x_n^t|y_n, x_{<n}^0) \propto p(y_n|x_n^t) p(x_n^t|x_{<n}^0) \tag{21}$$

for all the reverse time steps. Because

$$p(x_n^t|x_{<n}^0) = \int p(x_n^t|x_n^{t+1}, x_{<n}^0) p(x_n^{t+1}) dx_n^{t+1} \tag{22}$$

and

$$\int p(x_n^t|x_n^{t+1}, y_n, x_{<n}^0) p(x_n^{t+1}) dx_n^{t+1} = p(x_n^t|y_n, x_{<n}^0) \,, \tag{23}$$

$$= \frac{p(y_n|x_n^t) p(x_n^t|x_{<n}^0)}{p(y_n)} \,, \tag{24}$$

then we have

$$\int p(x_n^t|x_n^{t+1}, y_n, x_{<n}^0) p(x_n^{t+1}) dx_n^{t+1} = \frac{p(y_n|x_n^t)}{p(y_n)} \cdot \int p(x_n^t|x_n^{t+1}, x_{<n}^0) p(x_n^{t+1}) dx_n^{t+1}. \tag{25}$$

Therefore, we have

$$p(x_n^t|x_n^{t+1}, y_n, x_{<n}^0) = \frac{p(y_n|x_n^t) p(x_n^t|x_n^{t+1}, x_{<n}^0)}{p(y_n)} \,. \tag{26}$$

$p(y_n)$ is a constant for evidence. Then with gradient based method, the posterior $p(x_n^t|x_n^{t+1}, y_n, x_{<n}^0)$ is sampled from the likelihood $p(y_n|x_n^t)$ and the reverse process $p(x_n^t|x_n^{t+1}, x_{<n}^0)$,

C Likelihood function for k-space

The autocalibration signal (ACS) region are lines through the center of k-space, however, are fully sampled. The sensitivity of a coil is a spatial profile that describes the receiving field that induces signals in the coil. The simultaneous data acquisition, with each coil's sensitivity corresponding to a different subregion, leads to a complete image without aliasing artifacts.

 https://doi.org/10.52202/079017-4100

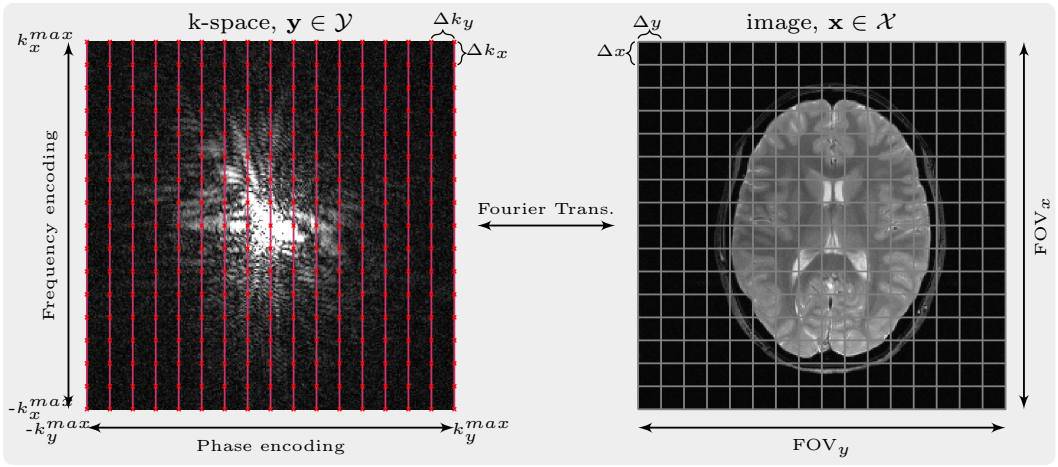


Figure 7: The relationship between k-space and image. The Nyquist theorem states that the sampling rate must be at least twice the highest frequency component in the signal.

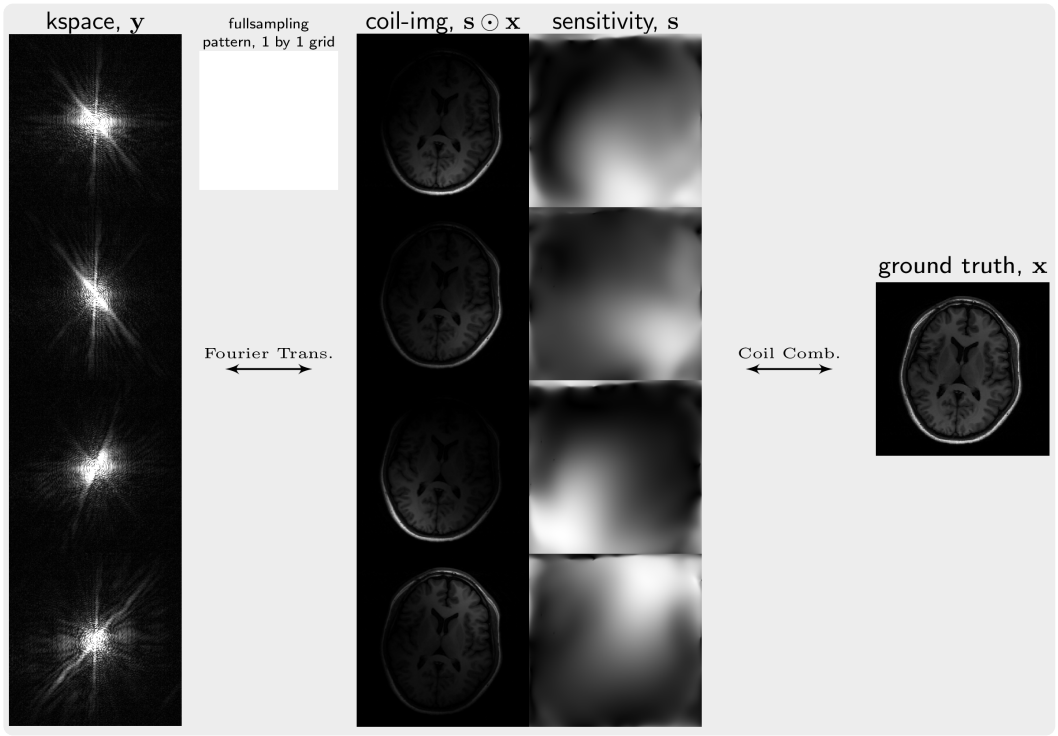


Figure 8: The signal detected by a coil is weighted by its local coil profile, which is called sensitivities and imposes weights on the signal intensity. Consequently, it causes dark and bright regions in coil images. The ground truth image is the combination of all coil images.

D Cardiac samples

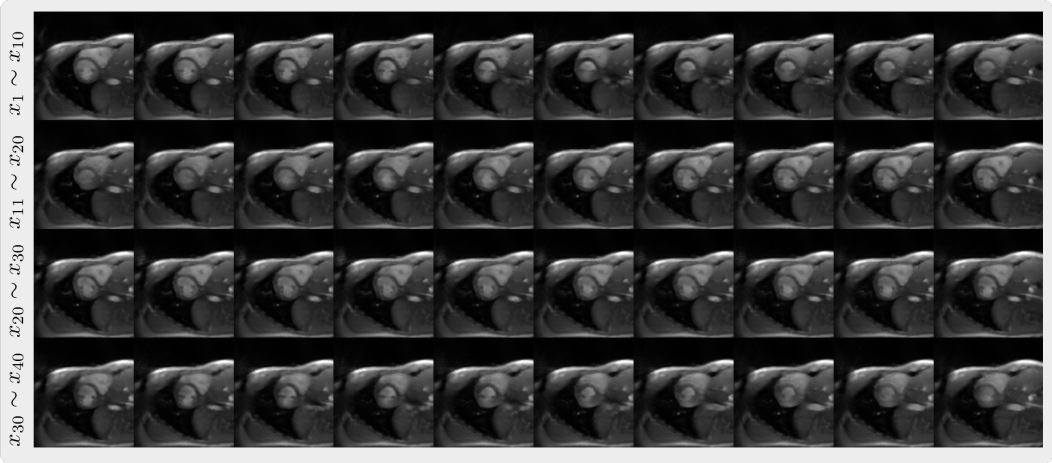


Figure 9: Prospective samples from the model trained on cardiac dataset.

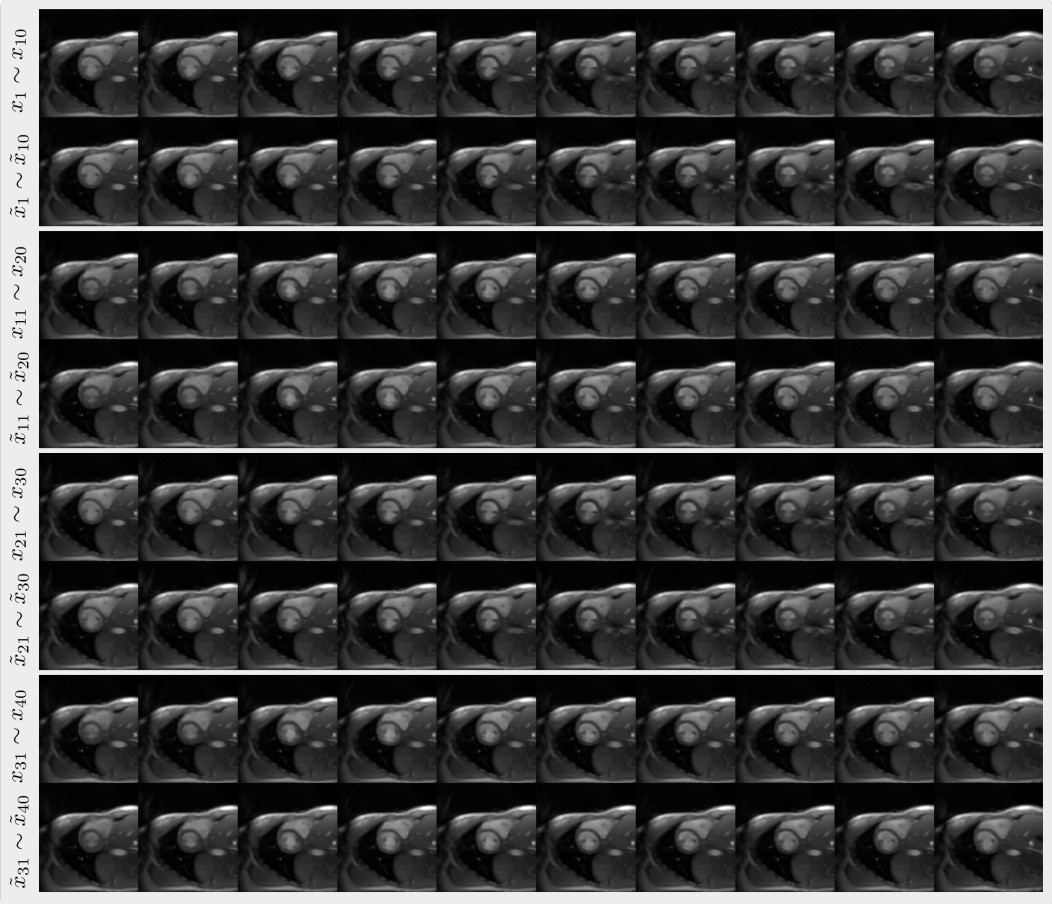
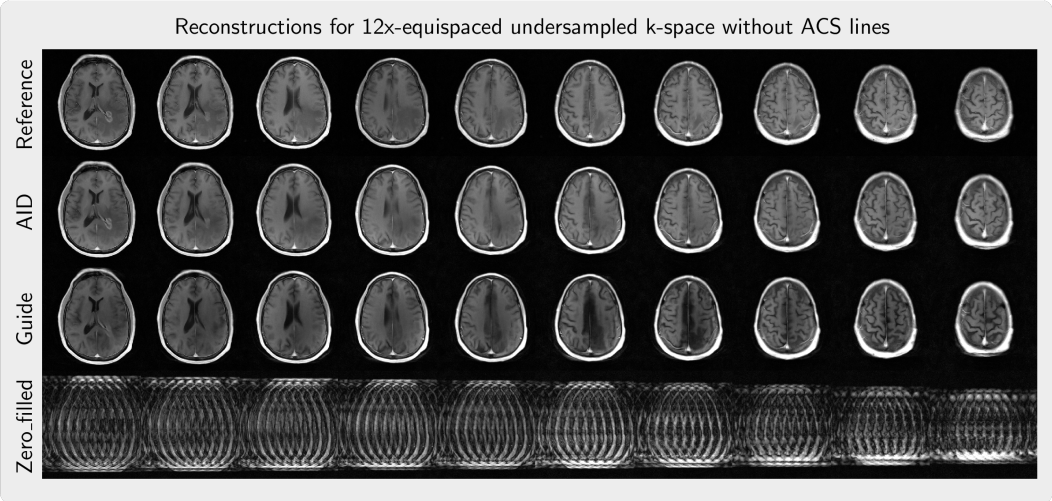
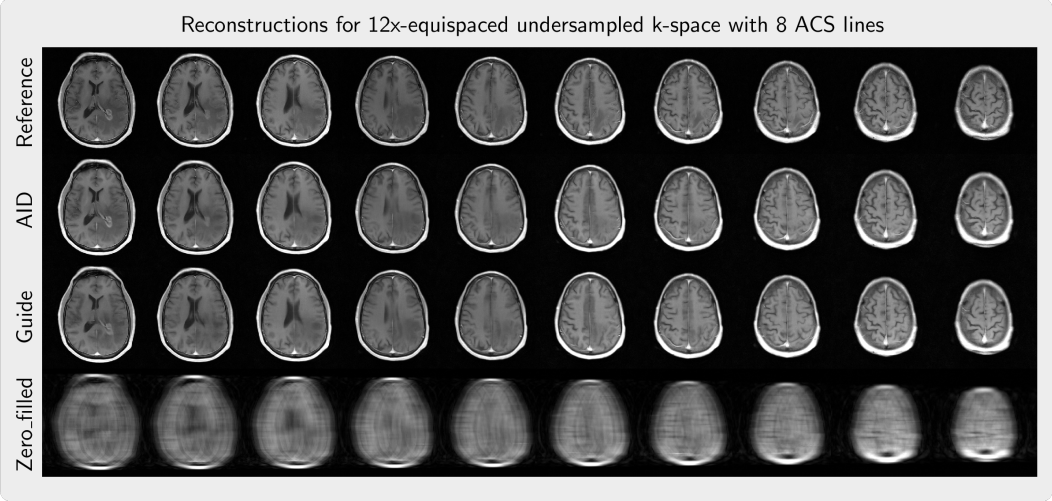


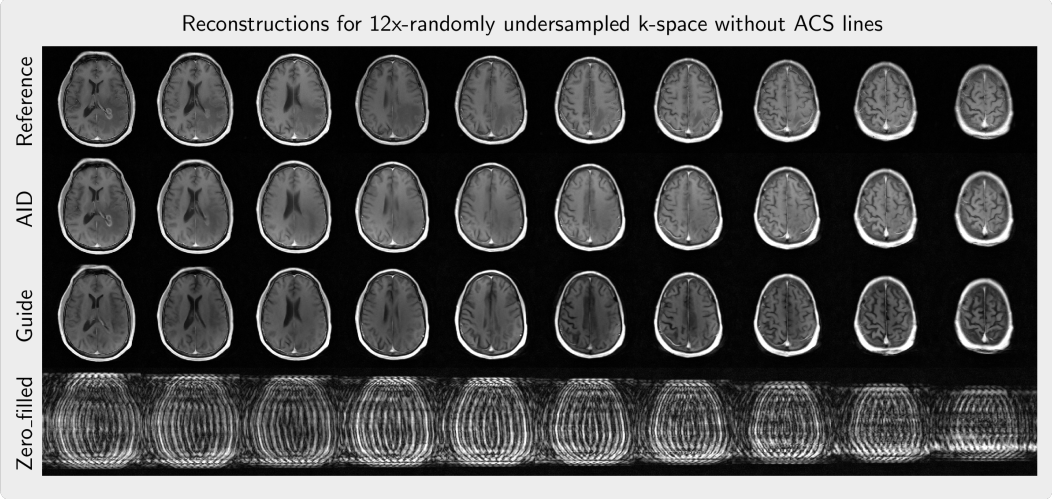
Figure 10: Retrospective samples from the model trained on cardiac dataset.

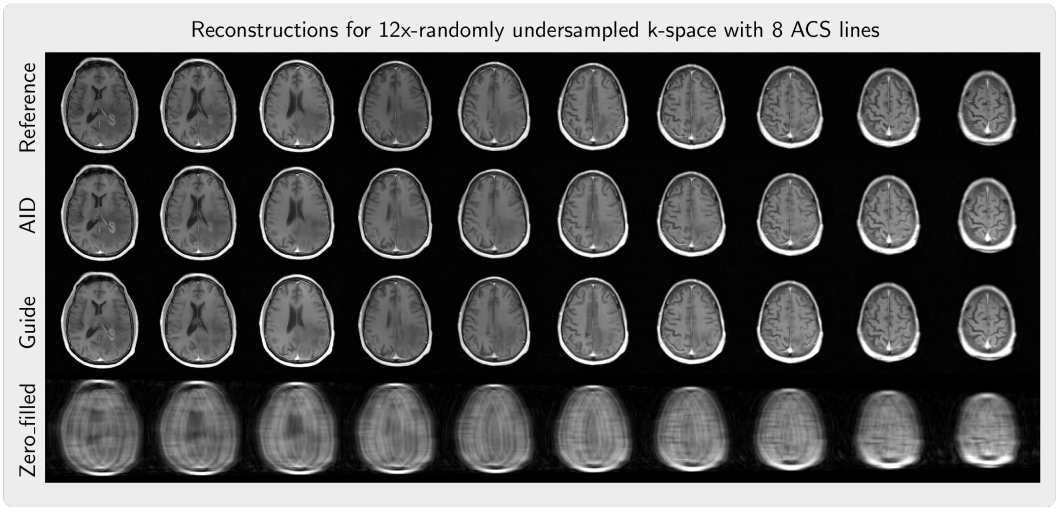
https://doi.org/10.52202/079017-4100

E Reconstruction from undersampled data









F Computation and model complexity

Table 1 presents the computation needed to train the AID models on four different datasets with different model complexities.

- Dataset: the name of the dataset (fastMRI, cardiac, ABIDE, UAV).
- Length: the length of the image sequence.
- Image size: the dimensions of the images in the dataset.
- Latent: the latent space representation used for the model, with options like VQVAE, Autoencoder-KL, or None.
- Two-stage: a boolean indicating whether a two-stage training process was used. Two-stage training is explained in the following section.
- Parameters: the number of parameters in the model.
- Train steps/s: training speed in steps per second.
- Inference (it/s): inference speed in iterations per second.

Table 1: Datasets and computational resources used to train the four different AID models.

Dataset	Length	Image size	Latent	Two-stage	Parameters	Train steps/s	Inference (it/s)
fastMRI	10	320	None	False	~139M	~1.31	~10.07
cardiac	42	256	VQVAE, 4x	False	~26M	~1.27	~20.20
ABIDE	46	128	None	True	~36M	~0.89	~4.10
UAV	70	256	Autoencoder-KL, 8x	True	~83M	~1.05	~39.60

G Two-stage training

We implemented a two-stage training process to improve training efficiency. In the first stage, we trained the U-net model. In the second stage, we trained the temporal-spatial conditioning block with the pre-trained U-net model frozen. By doing so, we are able to train an AID model on ABIDE dataset, where the image sequence has a dimension of 46×128×128 after preprocessing. The generated image squeence is shown in Figure 11.

H Natural image sequence generation

We trained an AID model on an UAV dataset in the latent space and generated images using the trained model. The generated images are displayed in Figure 12. The generated images demonstrate the effectiveness of the proposed method in learning sequentially coherent natural images generation. Each frame in Figure 2 shows an aerial view of a rural landscape with roads and/or a water pond.

https://doi.org/10.52202/079017-4100

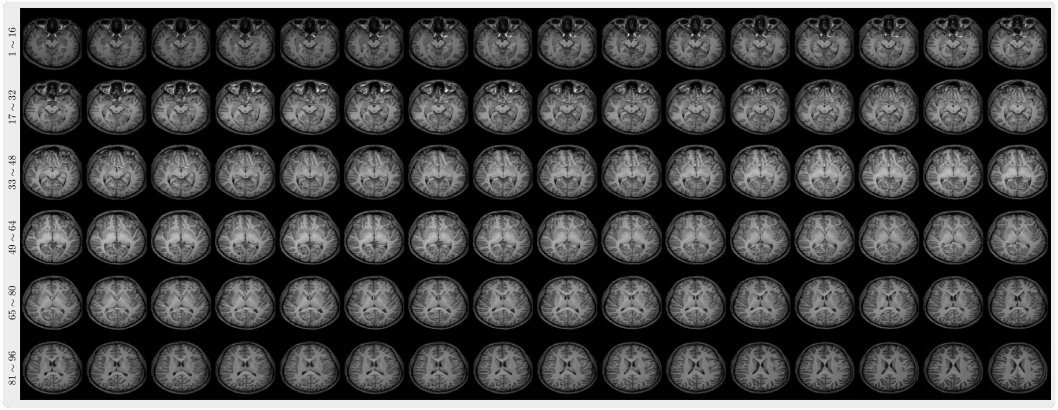


Figure 11: The generated ABIDE image sequence captures the changes in the brain structure.

Figure 12: The light changes in the water surface are captured in the generated UAV image sequence.

I Sample consistency along the temporal axis

Figure 13 shows the sample consistency along the temporal (or z) axis. Columns 1 and 2: Show sagittal and coronal views of a brain image sequence. These images appear to be medical scans with clearly stretched anatomical structures. Column 3: Displays the x-t plane of a cardiac image sequence. This displays the heart's activity over time and shows the diastolic and systolic phases from left to right. Columns 4 and 5: Show the x-t plane of a UAV image sequence, both generated and real. These images show the change in aerial views of a landscape over time. The generated x-t plane are generally consistent with the real x-t plane images but suffer from the striped artifacts.

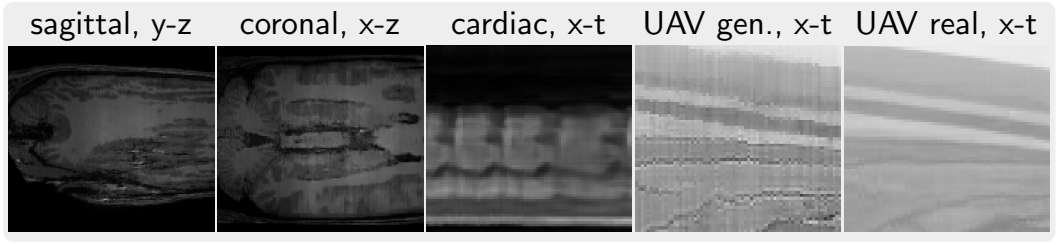


Figure 13: Temporal consistency of the images generated by AID models trained on different datasets. The first two columns show the sagittal and coronal view of brain image sequence. The x-t plane of cardiac image and UAV sequence are shown in the last three columns.

https://doi.org/10.52202/079017-4100

J VQVAE configuration for cardiac dataset

The VQVAE is trained on the cardiac dataset to generate the latent space for the training of the autoregressive diffusion model, using the official implementation[3]. The VQVAE is trained with the following configuration:

```
base_learning_rate: 4.5e-06
params:
  embed_dim: 3
  n_embed: 8192
  ddconfig:
    double_z: false
    z_channels: 3
    resolution: 256
    in_channels: 3
    out_ch: 3
    ch: 128
    ch_mult: [1, 2, 4]
    num_res_blocks: 2
    attn_resolutions: []
    dropout: 0.0
  lossconfig:
    target: losses.vqperceptual.VQLPIPSWithDiscriminator
    params:
      disc_conditional: false
      disc_in_channels: 3
      disc_start: 30001
      disc_weight: 0.8
      codebook_weight: 1.0
```

K 3D volume generation

To further improve the model's ability to generate 3D volumes, the position embedding is added to the third dimension - z of the volume. This allows the model trained on the ABIDE dataset to have better consistency along the z-axis.

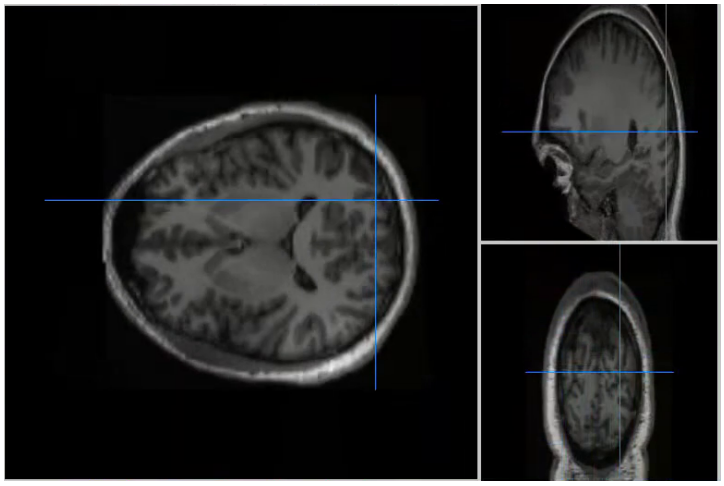


Figure 14: The generated 3D volumes from the model trained on the ABIDE dataset. The diffusion process is applied on the transverse plane (x-y) (c.f. the image on the left) and the autoregressive process is applied on the z-axis.

[3]https://github.com/CompVis/taming-transformers.git

https://doi.org/10.52202/079017-4100

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See the last paragraph of the introduction in Page 2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the last paragraph of discussion in Page 10.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See Appendix A and B

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides all the information needed to reproduce the main experimental results, include data, training details, algorithm details, and evaluation metrics in the section of experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-4100

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is available at `https://github.com/mrirecon/aid`.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides all the necessary details to understand the results in the section of experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bar in Figure 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 3.1

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

https://doi.org/10.52202/079017-4100

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: See **Privacy Issue** in the section of discussion.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See **Privacy Issue** in the section of discussion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release data or models that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets are properly credited and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not submit new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/079017-4100

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: This paper paper use publicly released fastMRI dataset and have been approved by the local IRB.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.