BIAS-STRUCTURE-CORRELATION TRADE OFF LEARNING FOR SELF-SUPERVISED DIFFUSION MRI DENOISING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatially correlated noise violates the input–target independence assumed by self-supervised diffusion MRI denoising. We identify a three-way trade-off among residual noise correlation, perturbation-induced structural distortion, and model-induced supervision bias, and propose a tensor-guided framework to balance them. By decomposing prediction error into conditional noise bias and structural mismatch, we derive an anisotropy-dependent bound on angular signal distortion under the diffusion tensor model. This bound motivates FA-guided spatio-angular shuffling that limits perturbations in highly anisotropic regions. To address residual noise dependence under restricted perturbations, we introduce a complementary tensor-reconstructed view that excludes the target direction from fitting, preventing direct target-noise reuse. Bidirectional auxiliary guidance and cross-view consistency combine the views to mitigate noise-related and model-induced biases. Experiments on a digital phantom and two in vivo datasets demonstrate improved DW image quality and preservation of diffusion-derived structures. Under 10% Rician noise, our method achieves 34.26 dB PSNR, outperforming the strongest baseline by 0.79 dB, and reduces FA estimation error by approximately 30% relative to the best competing method.

1 INTRODUCTION

Diffusion magnetic resonance imaging (dMRI) probes tissue microstructure through water diffusion, supporting white-matter analysis and quantitative biomarker estimation (Basser et al., 1994; Le Bihan, 2003; Le Bihan et al., 2006). However, acquisition constraints often yield low-SNR diffusion-weighted (DW) images, making denoising essential for reliable diffusion analysis. With clean references rarely available, self-supervised learning offers a practical approach to this problem.

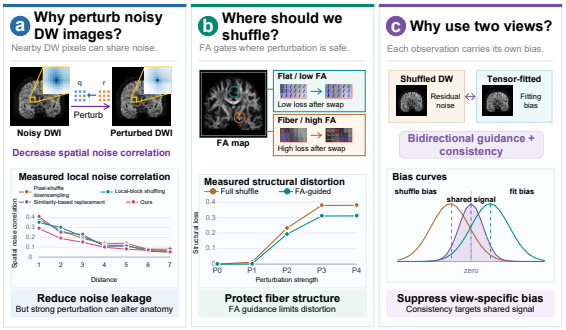


Figure 1: Motivation and core idea of the proposed method. (a) Perturbations reduce spatial noise correlation but may distort anatomy. (b) FA-guided shuffling limits fiber distortion. (c) Bidirectional guidance and consistency mitigate residual-noise and tensor-fitting biases.

A central challenge is that noisy observations do not necessarily provide unbiased supervision. Common self-supervised formulations assume zero-mean target noise sufficiently independent of the network input (Lehtinen et al., 2018; Batson & Royer, 2019). In dMRI, acquisition and reconstruction can introduce spatial noise correlation (Henriques et al., 2023; Wu et al., 2025), allowing neighboring voxels to leak target-correlated noise into blind-spot predictions (Krull et al., 2019). Perturbations such as downsampling (Zhou et al., 2020; Jang et al., 2023), shuffling (Chihaoui & Favaro, 2024), and replacement (Ma et al., 2025; Wu et al., 2025) reduce this dependence, but can also alter the clean signal that the network is expected to recover (Fig. 1(a)).

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

This tension is particularly important in dMRI because signal variation depends on both spatial location and diffusion direction. Cross-direction perturbations may preserve the signal in near-isotropic tissue while producing substantial mismatch in highly anisotropic regions. A uniform perturbation rule therefore cannot account for regional differences in structural sensitivity: strong perturbations risk distorting direction-dependent signals, whereas conservative perturbations retain noise dependence (Fig. 1(b)). The challenge is thus not merely to decorrelate noise, but to determine where perturbations preserve the signal correspondence needed for self-supervised prediction.

Even structurally adaptive perturbations leave a further supervision problem. Restricting perturbations protects sensitive structures but limits noise decorrelation. A model-derived view can provide complementary structural information, yet its accuracy is constrained by model mismatch. Consequently, supervision based on either perturbed observations or model-derived signals remains susceptible to bias (Fig. 1(c)). These limitations motivate a three-way trade-off among residual noise correlation, structural distortion, and model-induced supervision bias.

We address this trade-off with a tensor-guided self-supervised framework. Our analysis decomposes prediction error into conditional noise bias and structural mismatch, and derives an anisotropy-dependent bound on angular signal distortion under the diffusion tensor model. This motivates fractional anisotropy (FA)-guided spatio-angular shuffling that adapts perturbations to local structural sensitivity. To address the remaining supervision bias, we combine the shuffled view with a direction-excluded tensor-reconstructed view through bidirectional guidance and prediction consistency. Together, these designs balance noise decorrelation and structural preservation while reducing reliance on a single biased view. Our main contributions are:

1) We analyze self-supervised prediction error through conditional noise bias and structural mismatch, identifying a three-way trade-off among noise correlation, structural distortion, and supervision bias.
2) We derive an anisotropy-dependent angular distortion bound that motivates FA-guided spatio-angular shuffling for structurally adaptive noise decorrelation.
3) We develop complementary cross-view learning with direction-excluded tensor reconstruction to mitigate residual noise-related and model-induced biases.
4) Experiments on a digital phantom and two in vivo datasets demonstrate improved DW image quality and preservation of diffusion-derived structures and quantitative biomarkers.

2 Related Work

2.1 Self-Supervised Image Denoising

Self-supervised denoising learns from noisy observations without clean references. Noise2Noise (Lehtinen et al., 2018) uses independently corrupted observations of the same signal, while blind-spot and masking-based methods (Krull et al., 2019; Batson & Royer, 2019; Jiang et al., 2024) predict withheld pixels from their context. Their standard formulations assume zero-mean target noise sufficiently independent of the input. Spatially correlated noise can violate this assumption, allowing target-correlated noise to leak into predictions.

Existing methods mitigate this dependence through spatial separation or observation replacement. Pixel-shuffle downsampling (Zhou et al., 2020) increases pixel spacing within rearranged subimages, while AP-BSN (Lee et al., 2022) adopts asymmetric downsampling factors to balance decorrelation and detail preservation. PUCA (Jang et al., 2023) and related architectures (Li et al., 2025) incorporate unshuffle operations, whereas other methods suppress nearby observations (Li et al., 2023; Wang et al., 2023; Chen et al., 2024; Park et al., 2026). Alternatively, MASH (Chihaoui & Favaro, 2024) locally exchanges pixels with similar estimated clean intensities. Patch-Craft (Vaksman & Elad, 2023), Noise2Sim (Niu et al., 2023), and Pixel2Pixel (Ma et al., 2025) construct observations using patch- or pixel-level similarity. These strategies reduce noise dependence, but may sacrifice informative context or introduce signal mismatch through imperfect matching.

2.2 Diffusion MRI Denoising

dMRI denoising commonly exploits redundancy across spatial neighborhoods and diffusion volumes. PCA-based methods (Manjón et al., 2013; Ramos-Llordén et al., 2021; Mosso et al., 2022) represent local patches in low-rank subspaces and discard noise-dominated components.

Patch2Self (Fadnavis et al., 2020) and Patch2Self2 (Fadnavis et al., 2024) instead predict each diffusion volume from the others, avoiding explicit signal modeling. Their effectiveness nevertheless depends on sufficient local redundancy or cross-volume predictability and suitable noise assumptions.

Learned and physical priors provide complementary guidance. DDM^2 (Xiang et al., 2023) introduces a generative diffusion prior, whereas diffusion-specific methods (Tian et al., 2020; 2022) use model constraints or learned representations to preserve tissue information. UP2-MAE (Deng et al., 2026) combines a masked autoencoder with a DTI-based Noise2Noise branch through uncertainty-guided fusion and consistency regularization. Although these priors enrich the available supervision, model-derived signals remain vulnerable to approximation and model-mismatch errors. Simulation dataset under Self-supervised performance also depends critically on how training observations are constructed. Replace2Self (Wu et al., 2025) uses q-space block matching and consistency regularization to reduce noise dependence while preserving signal correspondence. However, because DW signals vary with both anatomy and diffusion direction, imperfect matching can introduce structural mismatch. More generally, strong perturbations improve noise decorrelation but distort direction-dependent structures, while conservative perturbations preserve structure at the cost of residual noise correlation. Model-derived guidance introduces an additional approximation bias. These competing effects motivate a unified formulation of conditional noise bias and clean-signal mismatch, enabling a principled balance among noise decorrelation, structural preservation, and model-based supervision.

3 METHOD

3.1 PROBLEM FORMULATION

Let $Y \in \mathbb{R}^{|\Omega| \times Q}$ denote a 4D diffusion MRI image, where Ω is the 3D spatial domain and Q is the number of diffusion directions. For voxel $i \in \Omega$ and direction $q \in \{1, \dots, Q\}$, the observed signal is modeled as $Y_{i,q} = S_{i,q} + N_{i,q}$, where $S_{i,q}$ is the clean diffusion signal and $N_{i,q}$ denotes noise.
In diffusion MRI, noise within each 3D volume is often spatially correlated, i.e., $\mathrm{Cov}(N_{i,q}, N_{j,q}) \neq 0$ for neighboring voxels $j \in \mathcal{N}(i)$, where $\mathcal{N}(i)$ is a local spatial neighborhood. By contrast, noises from different diffusion directions are approximately uncorrelated, i.e., $\mathrm{Cov}(N_{i,q}, N_{j,r}) \approx 0$ for $r \neq q$, where $r \in \{1, \dots, Q\}$ denotes another diffusion direction. A standard blind-spot denoiser predicts $Y_{i,q}$ from the same-direction spatial context $Z^{\mathrm{bs}}_{i,q} = \{Y_{j,q} : j \in \mathcal{N}(i), j \neq i\}$. Under the self-supervised squared loss, the optimal predictor is

$$f^*(Z^{\mathrm{bs}}_{i,q}) = \mathbb{E}[Y_{i,q} \mid Z^{\mathrm{bs}}_{i,q}] = \mathbb{E}[S_{i,q} \mid Z^{\mathrm{bs}}_{i,q}] + \mathbb{E}[N_{i,q} \mid Z^{\mathrm{bs}}_{i,q}]. \tag{1}$$

If the input noise were independent of the target noise, the second term would vanish. However, due to spatial noise correlation, $\mathbb{E}[N_{i,q} \mid Z^{\mathrm{bs}}_{i,q}] \neq 0$, leading to a noise-correlation bias in self-supervised denoising.
To mitigate this bias, we introduce a shuffling operator T_η to construct a blind-spot input $Z^\eta_{i,q} = T_\eta(Y; i, q)$ with $Y_{i,q} \notin Z^\eta_{i,q}$. Specifically, for each target pair (i, q), T_η selects a subset of neighboring voxels in $\mathcal{N}(i) \setminus \{i\}$ and replaces their same-direction observations $Y_{j,q}$ with cross-direction observations $Y_{j,r}$, where $r \neq q$ is sampled from the remaining diffusion directions. The parameter η controls the sampling rule of j and r. Since cross-direction noises are approximately uncorrelated, increasing such replacement reduces the residual correlation between the target noise $N_{i,q}$ and the input noise in $Z^\eta_{i,q}$. We denote this residual correlation by $R_{\mathrm{corr}}(\eta) = \mathbb{E}[\| \mathrm{Corr}(N_{i,q}, N(Z^\eta_{i,q}))\|^2]$, where $N(Z^\eta_{i,q})$ is the noise component of the shuffled input. Reducing $R_{\mathrm{corr}}(\eta)$ suppresses the conditional noise bias $\mathbb{E}[N_{i,q} \mid Z^\eta_{i,q}]$.
However, diffusion signals are direction-dependent, namely $S_{i,q} \neq S_{i,r}$ for $r \neq q$. Therefore, excessive shuffling may break direction-specific diffusion structure and make the context less informative for predicting $S_{i,q}$. For the shuffled input, the bias of the optimal self-supervised predictor can be written as

$$f^*(Z^\eta_{i,q}) - S_{i,q} = \underbrace{\mathbb{E}[N_{i,q} \mid Z^\eta_{i,q}]}_{\text{noise-correlation bias}} + \underbrace{\big(\mathbb{E}[S_{i,q} \mid Z^\eta_{i,q}] - S_{i,q}\big)}_{\text{structural bias}}. \tag{2}$$

This reveals a trade-off: stronger shuffling reduces noise-correlation bias but may increase structural bias by weakening angular diffusion consistency. We therefore learn a denoiser f_θ together with a shuffling strategy T_η to balance correlation reduction and structure preservation.

3.2 REGION-ADAPTIVE ANGULAR SHUFFLING FOR REDUCING STRUCTURE BIAS

Following the above trade-off, we instantiate the shuffling strategy T_η as a controlled spatio-angular operator (Fig. 2). For each target index (i, q), T_η samples a nearby spatial location $j \in \mathcal{N}(i) \setminus \{i\}$ and a diffusion direction $r \neq q$, where η controls the spatial range and angular sampling rule. The shuffled observation is $T_\eta(Y)_{i,q} = Y_{j,r}$, which can be decomposed as

$$T_\eta(Y)_{i,q} = S_{i,q} + \underbrace{(S_{j,r} - S_{i,q})}_{\mathcal{D}_{T_\eta}} + \underbrace{N_{j,r}}_{N_{T_\eta}}, \tag{3}$$

here, $\mathcal{D}_{T_\eta}$ denotes the clean-signal mismatch introduced by shuffling, while N_{T_η} is the replaced noise realization. Since $N_{j,r}$ comes from a different diffusion direction, it is less correlated with the target noise $N_{i,q}$, but the induced clean mismatch must be controlled.

We assume the clean diffusion signal is locally Lipschitz continuous in the joint spatio-angular domain. Then

$$\left|\mathcal{D}_{T_\eta}(i, q)\right| \leq L_{\mathrm{spa}}(i)\|j - i\|_2 + L_{\mathrm{ang}}(i)\, d_{\mathbb{S}^2}(\mathbf{g}_r, \mathbf{g}_q), \tag{4}$$

where $L_{\mathrm{spa}}(i)$ and $L_{\mathrm{ang}}(i)$ are the local spatial and angular Lipschitz constants, and $d_{\mathbb{S}^2}(\mathbf{g}_r, \mathbf{g}_q) = \arccos(|\langle \mathbf{g}_r, \mathbf{g}_q \rangle|)$ is the geodesic distance between unit gradient directions. Hence the structural distortion is jointly controlled by the spatial displacement, the angular separation, and the local signal smoothness.

Since the spatial shuffling radius is small by design, the dominant term in (4) is the angular one. To suppress angular distortion in highly anisotropic regions, we apply an FA-guided binary mask

$$M(i) = \begin{cases} 1, & \mathrm{FA}(i) < \tau_{\mathrm{FA}}, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

so that the masked observation is

$$\widetilde{Y}_{i,q} = M(i)\, Y_{j,r} + \big(1 - M(i)\big) Y_{i,q}. \tag{6}$$

The corresponding clean-signal distortion satisfies

$$\widetilde{\mathcal{D}}_{T_\eta}(i, q) = M(i)\, |S_{j,r} - S_{i,q}| \leq M(i)[L_{\mathrm{spa}}(i)\|j - i\|_2 + L_{\mathrm{ang}}(i)\, d_{\mathbb{S}^2}(\mathbf{g}_r, \mathbf{g}_q)]\,. \tag{7}$$

Under the diffusion tensor model, the angular Lipschitz constant is bounded by local anisotropy (proof can be found in Appendix A.1):

$$L_{\mathrm{ang}}(i) \leq C(i)\, \mathrm{FA}(i), \quad C(i) = \tfrac{2}{\sqrt{3}}\, b\, S_0(i)\, \|\boldsymbol{\lambda}(i)\|_2, \tag{8}$$

where b is the diffusion weighting, $S_0(i)$ the baseline signal, and $\boldsymbol{\lambda}(i)$ the tensor eigenvalue vector. In the masked region where $\mathrm{FA}(i) < \tau_{\mathrm{FA}}$, substituting $L_{\mathrm{ang}}(i) \leq C(i)\tau_{\mathrm{FA}}$ into (7) gives

$$\widetilde{\mathcal{D}}_{T_\eta}(i, q) \leq M(i)[L_{\mathrm{spa}}(i)\|j - i\|_2 + C(i)\, \tau_{\mathrm{FA}}\, d_{\mathbb{S}^2}(\mathbf{g}_r, \mathbf{g}_q)]. \tag{9}$$

Therefore, restricting shuffling to low-FA regions tightens the angular mismatch bound and limits the structural distortion introduced by T_η.

3.3 CROSS-VIEW AUXILIARY GUIDANCE FOR NOISE BIAS REDUCTION

As shown in Fig. 2, we construct the shuffled noisy view y^b and a target-conditioned tensor-fitted auxiliary view $y^{a,p}$. For location $p = (i, q)$, let $r_p \neq q$ be the direction selected by T_η, so that $y^b_p = Y_{j,r_p}$ in a shuffled region. The tensor fit excludes both the current direction and the selected source direction. Specifically, define

$$\mathcal{E}_p = \begin{cases} \{q, r_p\}, & M(i) = 1, \\ \{q\}, & M(i) = 0. \end{cases} \tag{10}$$

For every spatial location ℓ, fit the tensor using only directions outside $\mathcal{E}_p$ and reconstruct the target direction as

$$\widehat{\mathbf{D}}_\ell^{(-\mathcal{E}_p)} = \arg\min_{\mathbf{D} \succeq 0} \sum_{k \notin \mathcal{E}_p} \left[\log\left(\frac{Y_{\ell,k}}{S_{0,\ell}}\right) + b_k \mathbf{g}_k^\top \mathbf{D} \mathbf{g}_k\right]^2, \quad y^{a,p}_{\ell,q} = S_{0,\ell} \exp\Big(-b_q \mathbf{g}_q^\top \widehat{\mathbf{D}}_\ell^{(-\mathcal{E}_p)} \mathbf{g}_q\Big)\,. \tag{11}$$

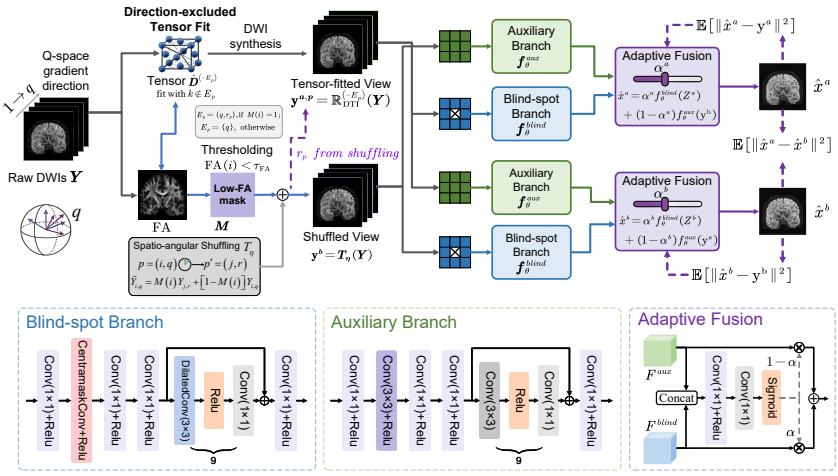


Figure 2: Overview of the proposed framework. It comprises FA-guided spatio-angular shuffling and a dual-branch cross-view denoiser. For each target, the tensor-fitted auxiliary view excludes the current direction and, where shuffling is applied, the selected source direction. Detailed network architecture description can be found in Appendix A.2

.

Below, y^a abbreviates the target-conditioned auxiliary view $y^{a,p}$. A dual-branch network comprises a blind-spot branch $f_\theta^{\mathrm{blind}}$ and an auxiliary branch f_θ^{aux}. Let Z_p^a and Z_p^b denote the center-excluded blind-spot neighborhoods of y^a and y^b. The network is applied bidirectionally:

$$\hat{x}_p^b = \alpha_p^b f_\theta^{\mathrm{blind}}(Z_p^b) + (1-\alpha_p^b) f_\theta^{\mathrm{aux}}(y^a)_p, \quad \hat{x}_p^a = \alpha_p^a f_\theta^{\mathrm{blind}}(Z_p^a) + (1-\alpha_p^a) f_\theta^{\mathrm{aux}}(y^b)_p, \tag{12}$$

where $\alpha_p^a, \alpha_p^b \in [0,1]$ are adaptive fusion weights. The predictions $\hat{x}^b$ and $\hat{x}^a$ are supervised by y^b and y^a, respectively, under the objective

$$\mathcal{L} = \mathbb{E}\left[\|\hat{x}^b - y^b\|^2\right] + w_a \mathbb{E}\left[\|\hat{x}^a - y^a\|^2\right] \quad + \lambda \mathbb{E}\left[\|\hat{x}^b - \hat{x}^a\|^2\right], \tag{13}$$

where w_a balances the two reconstruction losses and λ controls bidirectional consistency.
The direction-excluded fit removes the direct route from the shuffled target to the auxiliary tensor reconstruction. In particular, $y^a = h_p(Y_{-\mathcal{E}_p})$, where $Y_{-\mathcal{E}_p} = \{Y_{\cdot,k} : k \notin \mathcal{E}_p\}$. Excluding the full source direction r_p also removes its spatially correlated measurements from the auxiliary branch's input. For a fixed mask and selected source direction, $\partial y^a / \partial Y_{\cdot,d} = 0$ for each $d \in \mathcal{E}_p$. The center value $y_{i,q}^a$ therefore cannot directly copy the noisy target; it is predicted from the remaining directions. Approximate cross-direction independence limits noise transfer through those remaining measurements, but does not by itself remove magnitude-noise bias. We represent the residual conditional bias as

$$\delta_p^{a\to b} := \mathbb{E}[n_p^b \mid y^a] = \mathbb{E}[n_p^b \mid h_p(Y_{-\mathcal{E}_p})], \qquad |\delta_p^{a\to b}| \le \varepsilon_p |B_{\mathrm{BSN}}^b(p)|, \qquad 0 \le \varepsilon_p < 1, \tag{14}$$

where the bound is an assumption about the remaining dependence, rather than a consequence of direction exclusion alone. The term $\delta_p^{a\to b}$ can include residual magnitude-noise bias and dependence introduced by shared preprocessing.
For a standard blind-spot network targeting y^b, the MSE-optimal estimator satisfies

$$f_{\mathrm{BSN}}^*(Z_p^b) = \mathbb{E}[x_p \mid Z_p^b] + \underbrace{\mathbb{E}[n_p^b \mid Z_p^b]}_{B_{\mathrm{BSN}}^b(p)}. \tag{15}$$

In our dual-branch formulation, the bias with respect to n_p^b is

$$B_{\mathrm{dual}}^b(p) = \alpha_p^b \mathbb{E}[n_p^b \mid Z_p^b] + (1-\alpha_p^b)\mathbb{E}[n_p^b \mid y^a]. \tag{16}$$

Substituting Eq. (14) gives

$$B_{\mathrm{dual}}^b(p) = \alpha_p^b B_{\mathrm{BSN}}^b(p) + (1-\alpha_p^b)\delta_p^{a\to b}, \quad |B_{\mathrm{dual}}^b(p)| \le [\alpha_p^b + (1-\alpha_p^b)\varepsilon_p] |B_{\mathrm{BSN}}^b(p)|. \tag{17}$$

Under the stated bound, $\alpha_p^b < 1$ yields an upper bound below $|B_{\mathrm{BSN}}^b(p)|$. Thus auxiliary guidance can suppress noise bias without requiring a zero conditional noise mean. Symmetrically, $B_{\mathrm{dual}}^a(p) \approx \alpha_p^a \mathbb{E}[n_p^a \mid Z_p^a]$. Because y^a has higher SNR than y^b, $|\mathbb{E}[n_p^a \mid Z_p^a]| < |\mathbb{E}[n_p^b \mid Z_p^b]|$, so $|B_{\mathrm{dual}}^a| < |B_{\mathrm{dual}}^b|$ in general.
However, y^a may carry structural degradation. Writing $y^a = x + s^a + n^a$, the bias becomes

$$B_{\mathrm{dual}}^b(p) \approx \alpha_p^b \mathbb{E}[n_p^b \mid Z_p^b] + (1 - \alpha_p^b)\left(\delta_p^{a \to b} + \mathbb{E}[s_p^a \mid y^a]\right), \tag{18}$$

so the structural leakage from the auxiliary view must be controlled, this is the role of the bidirectional consistency term.
Denoting the pre-consistency conditional estimates as $\mu_v = x + B_{\mathrm{dual}}^v$ ($v \in \{a, b\}$), the per-pixel regularized objective is

$$\ell = (\hat{x}^b - \mu_b)^2 + w_a(\hat{x}^a - \mu_a)^2 + \lambda(\hat{x}^b - \hat{x}^a)^2. \tag{19}$$

Solving this quadratic system yields

$$\hat{x}^{b,*} = \frac{(w_a + \lambda)\mu_b + \lambda w_a \mu_a}{w_a + \lambda(1 + w_a)}, \tag{20}$$

and the resulting bias of branch b is

$$B_{\mathrm{con}}^b = (1 - \rho_b)B_{\mathrm{dual}}^b + \rho_b B_{\mathrm{dual}}^a, \quad \rho_b = \frac{\lambda w_a}{w_a + \lambda(1 + w_a)}. \tag{21}$$

Thus consistency regularization interpolates branch b's estimate toward the less biased branch a. When $|B_{\mathrm{dual}}^a| < |B_{\mathrm{dual}}^b|$, this reduces the final noise bias without imposing a hard constraint that could over-smooth view-specific structures.

4 Experimental Setup

Datasets. We evaluate on one digital phantom and two in vivo datasets. The Phantomas phantom (Caruyer et al., 2014) provides clean DW images with 64 diffusion directions at $b = 1000\ \mathrm{s/mm^2}$. We separately simulate Rician and noncentral Chi (NC-Chi) noise at 6%, 8% and 10% of the maximum signal intensity, using four coils for NC-Chi noise. The HCP dataset (Van Essen et al., 2013) contains 90 directions per shell at $b = \{1000, 2000, 3000\}\ \mathrm{s/mm^2}$; Sherbrooke 3-Shell (Garyfallidis et al., 2014) contains 64 directions per shell at $b = \{1000, 2000, 3500\}\ \mathrm{s/mm^2}$. The detailed noise simulation and data description are provided in Appendix A.3.

Evaluation metrics. To quantify denoising performance, we report peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS). We further use the average angular error (AAE) to evaluate fiber-orientation recovery. To examine the trade-off between noise decorrelation and structural distortion, we additionally report the mean spatial noise correlation ρ_{mean}, mean structure distortion $\overline{M}$, and noise expectation bias (NEB). Their definitions and implementation details are provided in Appendix A.4.

Implementation and baselines. We implement our method in PyTorch and train for 200 epochs using Adam with a learning rate of 10^{-4} and a batch size of 32 on an NVIDIA A6000 GPU. We compare against MPPCA, Patch2Self, SAN2N, Replace2Self, and UP2-MAE on the same test data. Baseline configurations are detailed in Appendix A.5.

5 Results and Analysis

5.1 Results on Simulation Dataset

Fig. 3 presents a visual comparison of DW image denoising results under 10% Rician noise; the results for other noise levels and NC-Chi noise are given in Appendix A.6. Our method produces cleaner images with fewer residual artifacts while better preserving fine anatomical details and structural integrity in zoomed-in views. As reported in Table 1, it achieves the highest PSNR and SSIM and the lowest LPIPS among the compared methods. The fiber tractography (FT) and orientation (FO) results (Fig. 3, rows 4-6) further confirm the downstream benefit. SAN2N and Patch2Self suffer from evident fiber loss and incomplete bundles, while Replace2Self hallucinates spurious fibers in homogeneous regions. MPPCA and UP2-MAE recover comparatively complete bundles

but yield less regular and coherent streamlines. Our method best matches the ground-truth orientations, achieving the lowest AAE of 3.83 ± 2.97, a 4.96% reduction over the second-best MP-PCA (4.03 ± 2.97). The bottom rows of Fig. 3 show the FA/MD maps and residuals. SAN2N and Patch2Self deviate noticeably from the ground truth; Replace2Self, MPPCA, and UP2-MAE preserve the main structures but retain visible residuals near fiber regions, tissue boundaries, and central or peripheral areas. Our method produces the most faithful FA and MD maps with the fewest artifacts, attaining the lowest FA and MD MAE (0.0514 and 0.0276, Table 1).

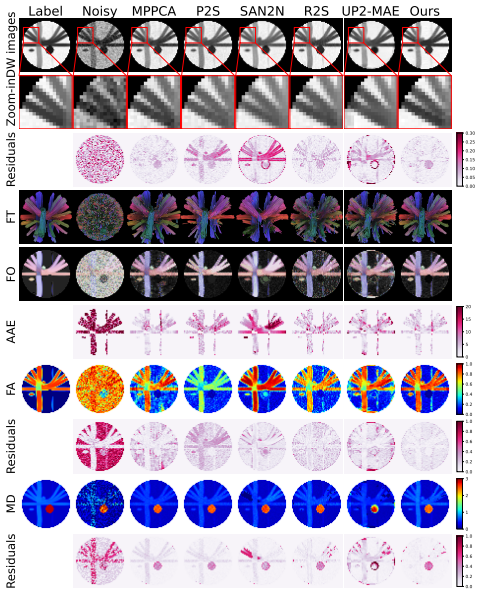


Figure 3: Qualitative comparison of different denoising methods on the Simulation dataset (Rician noise level = 10%).

Table 1: Quantitative comparison of DWI quality and DTI parameter estimation under 10% Rician noise. Bold indicates the best result.

	Method	PSNR↑	SSIM↑	LPIPS↓
DWI metrics	Noisy	21.10 ± 0.16	0.762 ± 0.0157	0.073 ± 0.0105
	MPPCA	33.47 ± 0.46	0.952 ± 0.0029	0.002 ± 0.0012
	P2S	31.07 ± 1.62	0.951 ± 0.0051	0.008 ± 0.0035
	SAN2N	27.45 ± 1.69	0.950 ± 0.0066	0.015 ± 0.0079
	R2S	32.29 ± 0.75	0.956 ± 0.0022	0.004 ± 0.0023
	UP2-MAE	28.01 ± 0.46	0.948 ± 0.0031	0.005 ± 0.0020
	Ours	$\mathbf{34.26 \pm 0.37}$	$\mathbf{0.960 \pm 0.0036}$	$\mathbf{0.002 \pm 0.0010}$
	Method	FA MAE↓	MD MAE↓	AAE (°)↓
DTI metrics	Noisy	0.279 ± 0.330	0.154 ± 0.222	22.44 ± 16.60
	MPPCA	0.091 ± 0.117	0.040 ± 0.076	4.03 ± 2.97
	P2S	0.120 ± 0.112	0.034 ± 0.067	5.97 ± 3.77
	SAN2N	0.073 ± 0.087	0.048 ± 0.115	8.27 ± 5.03
	R2S	0.080 ± 0.087	0.035 ± 0.084	5.76 ± 4.41
	UP2-MAE	0.088 ± 0.125	0.077 ± 0.173	6.19 ± 6.37
	Ours	$\mathbf{0.051 \pm 0.058}$	$\mathbf{0.028 \pm 0.083}$	$\mathbf{3.83 \pm 2.97}$

P2S: Patch2Self; R2S: Replace2Self.

5.2 Results on Real Acquisitions

Fig. 4 compares qualitative results on HCP and Sherbrooke 3-Shell. On HCP (a), MPPCA and SAN2N retain noise, with SAN2N also suppressing anatomical detail, while Patch2Self, Replace2Self, and UP2-MAE exhibit oversmoothing. These errors coincide with sparse or disorganized tractography, orientation deviations, and missing fibers. Our method better preserves anatomical detail and coherent fiber bundles, with cleaner low-FA regions and retained high-FA structure; MD maps are broadly comparable across methods. On Sherbrooke 3-Shell (b), fewer diffusion directions and stronger noise make denoising more challenging. Replace2Self suppresses anatomical signals and introduces FA artifacts, SAN2N and UP2-MAE retain substantial noise, and MPPCA and Patch2Self oversmooth high-FA regions. Our method better balances noise suppression and structural preservation, yielding more coherent tractography and better-defined FA maps.

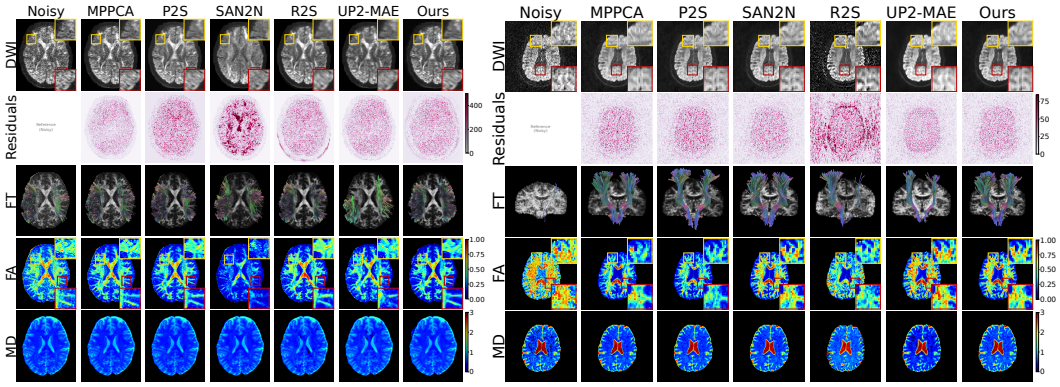
5.3 Ablation Results and Analysis

5.3.1 Effectiveness of the FA-guided mask

We evaluate the proposed FA-guided masking against four baselines: Zero (fully shuffled input), One (noisy observation retained), and spatially agnostic Random and Patch schemes. As shown in Table 2, FA-guided masking consistently achieves the best performance under Rician noise (10%), attaining the highest PSNR (34.26 dB) and SSIM (0.961) alongside the lowest LPIPS (0.002). This demonstrates that FA-derived structural priors yield a more effective masking configuration, preserving anatomically relevant information while mitigating the noise spatial correlation.

Table 2: Mask ablation under 10% Rician noise.

Strategy	PSNR↑	SSIM↑	LPIPS↓
Zero	33.27 ± 0.36	0.957 ± 0.004	0.004 ± 0.001
One	32.43 ± 0.44	0.954 ± 0.004	0.004 ± 0.002
Random	33.46 ± 0.51	0.957 ± 0.004	0.003 ± 0.002
Patch	32.46 ± 0.32	0.954 ± 0.004	0.004 ± 0.002
FA	$\mathbf{34.26 \pm 0.37}$	$\mathbf{0.961 \pm 0.004}$	$\mathbf{0.002 \pm 0.001}$



(a) HCP ($b = 3000\ \mathrm{s/mm^2}$). (b) Sherbrooke 3-Shell ($b = 1000\ \mathrm{s/mm^2}$).

Figure 4: Qualitative comparison of denoising methods on two real dMRI datasets.

5.3.2 Effectiveness of Cross-View Auxiliary Guidance and Consistency Regularization

Fig. 5 compares the noisy input and four configurations: DTI-Recon synthesizes DW images from tensors fitted to noisy data; BSN uses shuffling-based blind-spot denoising; Dual adds auxiliary DTI-reconstruction guidance; and Full enforces mutual consistency between the two cross-guided estimates. Quantitatively, Full achieves the best DW-image metrics and the lowest FA MAE and AAE, while DTI-Recon attains the lowest MD MAE. Qualitatively, DTI-Recon retains substantial background noise, while BSN disrupts fiber structures. Dual improves structural preservation but leaves boundary artifacts. Full better preserves fiber structures and suppresses background noise, supporting the complementary benefits of cross-view guidance and consistency regularization.

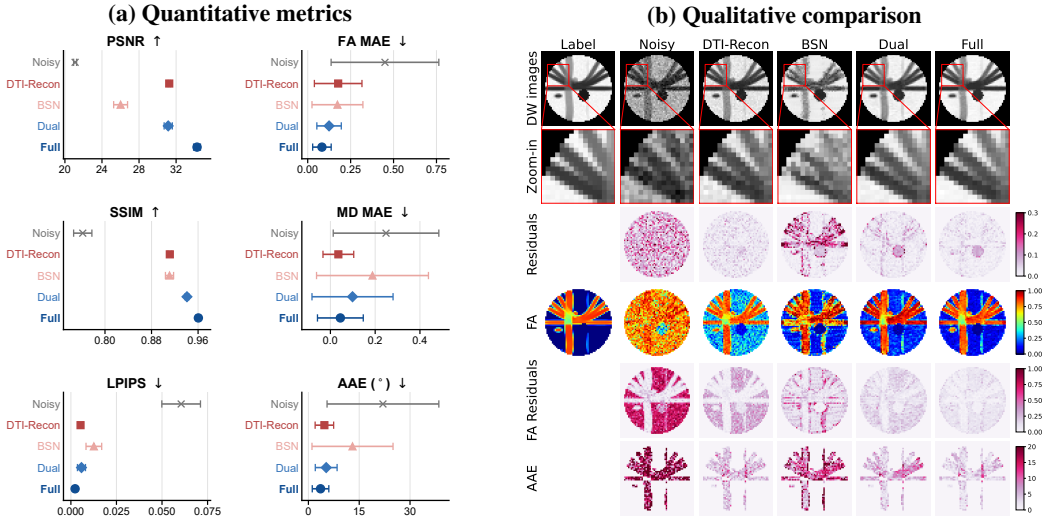
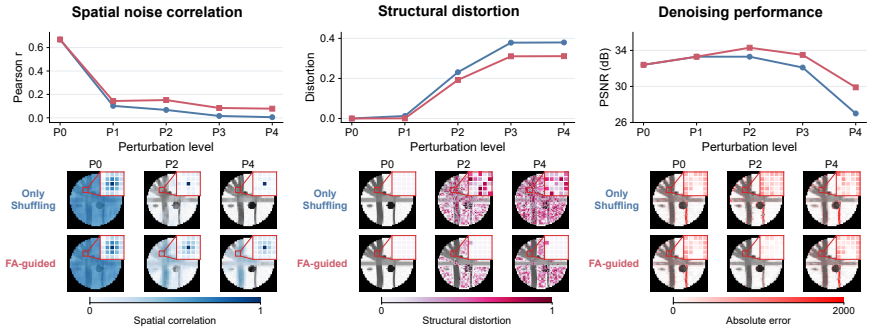


Figure 5: Quantitative and qualitative ablation of cross-view auxiliary guidance and consistency regularization under 10% Rician noise. The left panel shows means with standard-deviation error bars; arrows indicate the preferred metric direction.

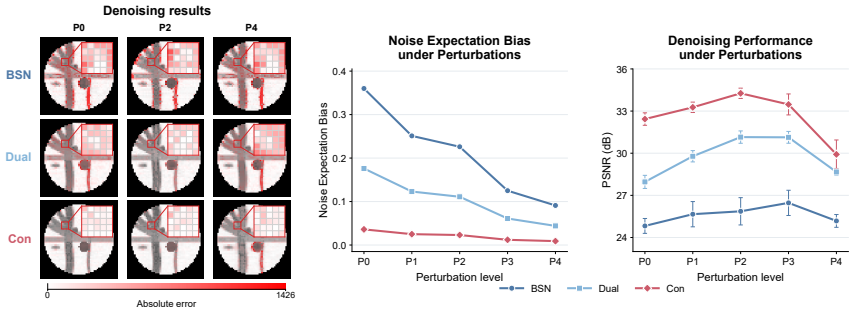
5.3.3 Analysis of Bias-Structure-Correlation Trade-off Learning

Fig. 6(a) examines the trade-off among noise decorrelation, structural preservation, and denoising performance. As the perturbation increases from P0 to P4, both strategies reduce spatial noise correlation, with random shuffling achieving stronger decorrelation. FA-guided shuffling retains higher correlation because it suppresses perturbations in structurally sensitive, high-FA regions. The spatial correlation maps further show that FA guidance primarily decorrelates homogeneous regions while preserving directional dependencies in anisotropic structures (blue region at P4). This selective per-

turbation produces substantially less structural distortion, as evidenced by both the distortion curves and spatial maps, where darker magenta indicates greater distortion. FA-guided shuffling consequently yields lower residuals and higher PSNR at moderate and strong perturbation levels. Performance peaks near P2, where sufficient noise decorrelation is achieved without excessive structural damage. By contrast, strong random shuffling at P4 introduces substantial structural bias and causes a sharp PSNR drop. These results show that stronger decorrelation alone is insufficient: by retaining limited correlation in sensitive regions, FA-guided shuffling better preserves structure and achieves a more favorable denoising trade-off. Fig. 6(b) shows that consistency supervision yields the small-



(a) Comparing noise decorrelation, structural distortion, and denoising performance with and without FA guidance.



(b) Noise expectation bias vs performance obtained with different supervisions.

Figure 6: Analysis of the trade-off among noise decorrelation, structural preservation, and denoising bias.

est absolute residuals and the highest PSNR across perturbation levels. It also achieves the lowest noise expectation bias (NEB), indicating that the constructed supervision better satisfies the conditional zero-mean assumption of self-supervised denoising. Although stronger perturbations further reduce NEB, performance peaks near P2 and then declines due to increasing structural mismatch. Consistency supervision therefore provides the best balance between supervision bias and structural preservation.

6 Conclusion

We studied self-supervised dMRI denoising through the trade-off among residual noise correlation, structural distortion, and model-induced supervision bias. Our analysis decomposes the error induced by perturbed observations into conditional noise bias and structural mismatch, and establishes an anisotropy-dependent bound on angular signal distortion under the diffusion tensor model. These results motivate FA-guided spatio-angular shuffling and complementary cross-view supervision using direction-excluded tensor reconstruction. Experiments on a digital phantom and two in vivo datasets demonstrate improved denoising and diffusion-derived structural fidelity, supporting the importance of balancing noise decorrelation with signal preservation.

Limitations and future work. Our angular distortion bound is derived under the diffusion tensor model and does not fully characterize non-tensor diffusion behavior, such as crossing fibers or high-b signal deviations. Extending the analysis to richer diffusion models would clarify the scope of anisotropy-guided perturbations. In addition, perturbation strength is selected empirically; adapting it to local noise dependence and structural sensitivity is a promising direction.

REFERENCES

Peter J Basser, James Mattiello, and Denis LeBihan. Mr diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267, 1994.

Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International conference on machine learning*, pp. 524–533. PMLR, 2019.

Emmanuel Caruyer, Alessandro Daducci, Maxime Descoteaux, Jean-Christophe Houde, Jean-Philippe Thiran, and Ragini Verma. Phantomas: a flexible software library to simulate diffusion mr phantoms. In *The International Society for Magnetic Resonance in Medicine (ISMRM)*, 2014.

Shiyan Chen, Jiyuan Zhang, Zhaofei Yu, and Tiejun Huang. Exploring efficient asymmetric blind-spots for self-supervised denoising in real-world scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2814–2823, 2024.

Hamadi Chihaoui and Paolo Favaro. Masked and shuffled blind spot denoising for real-world images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3025–3034, 2024.

Zeyu Deng, Lihui Wang, Xi Tao, Qijian Chen, Ying Cao, Yingfeng Ou, et al. Uncertainty-propelled physics-mae fusion for self-supervised diffusion-weighted image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3578–3586, 2026.

Shreyas Fadnavis, Joshua Batson, and Eleftherios Garyfallidis. Patch2self: Denoising diffusion mri with self-supervised learning. *Advances in neural information processing systems*, 33:16293–16303, 2020.

Shreyas Fadnavis, Agniva Chowdhury, Joshua Batson, Petros Drineas, and Eleftherios Garyfallidis. Patch2self2: Self-supervised denoising on coresets via matrix sketching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27641–27651, 2024.

Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, and Ian Nimmo-Smith. Dipy, a library for the analysis of diffusion mri data. *Frontiers in Neuroinformatics*, Volume 8 - 2014, 2014. ISSN 1662-5196. doi: 10.3389/fninf.2014.00008. URL `https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2014.00008`.

Rafael Neto Henriques, Andrada Ianuş, Lisa Novello, Jorge Jovicich, Sune N Jespersen, and Noam Shemesh. Efficient pca denoising of spatially correlated redundant mri data. *Imaging Neuroscience*, 1:imag–1–00049, 12 2023. ISSN 2837-6056. doi: 10.1162/imag_a_00049. URL `https://doi.org/10.1162/imag_a_00049`.

Hyemi Jang, Junsung Park, Dahuin Jung, Jaihyun Lew, Ho Bae, and Sungroh Yoon. Puca: Patch-unshuffle and channel attention for enhanced self-supervised image denoising. *Advances in Neural Information Processing Systems*, 36:19217–19229, 2023.

Haotian Jiang, Shu Zhang, Xuyun Wen, Hui Cui, Jun Lu, Islem Rekik, Jiquan Ma, and Geng Chen. Self-supervised denoising of diffusion mri data via spatio-angular noise2noise. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2129–2137, 2019.

Denis Le Bihan. Looking into the functional architecture of the brain with diffusion mri. *Nature reviews neuroscience*, 4(6):469–480, 2003.

Denis Le Bihan, Cyril Poupon, Alexis Amadon, and Franck Lethimonnier. Artifacts and pitfalls in diffusion mri. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 24(3):478–488, 2006.

Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Ap-bsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17725–17734, 2022.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2965–2974. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/lehtinen18a.html.

Junyi Li, Zhilu Zhang, Xiaoyu Liu, Chaoyu Feng, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Spatially adaptive self-supervised learning for real-world image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9914–9924, 2023.

Junyi Li, Zhilu Zhang, and Wangmeng Zuo. Rethinking transformer-based blind-spot network for self-supervised image denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4788–4796, 2025.

Qing Ma, Junjun Jiang, Xiong Zhou, Pengwei Liang, Xianming Liu, and Jiayi Ma. Pixel2pixel: A pixelwise approach for zero-shot single image denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4614–4629, 2025. doi: 10.1109/TPAMI.2025.3546870.

José V Manjón, Pierrick Coupé, Luis Concha, Antonio Buades, D Louis Collins, and Montserrat Robles. Diffusion weighted image denoising using overcomplete local pca. *PloS one*, 8(9): e73021, 2013.

Jessie Mosso, Dunja Simicic, Kadir Şimşek, Roland Kreis, Cristina Cudalbu, and Ileana O Jelescu. Mp-pca denoising for diffusion mrs data: promises and pitfalls. *NeuroImage*, 263:119634, 2022.

Chuang Niu, Mengzhou Li, Fenglei Fan, Weiwen Wu, Xiaodong Guo, Qing Lyu, and Ge Wang. Noise suppression with similarity-based self-supervised deep learning. *IEEE Transactions on Medical Imaging*, 42(6):1590–1602, 2023. doi: 10.1109/TMI.2022.3231428.

Junyoung Park, Youngjin Oh, and Nam Ik Cho. Tm-bsn: Triangular-masked blind-spot network for real-world self-supervised image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29877–29886, 2026.

Gabriel Ramos-Llordén, Gonzalo Vegas-Sánchez-Ferrero, Congyu Liao, Carl-Fredrik Westin, Kawin Setsompop, and Yogesh Rathi. Snr-enhanced diffusion mri with structure-preserving low-rank denoising in reproducing kernel hilbert spaces. *Magnetic resonance in medicine*, 86(3): 1614–1632, 2021.

Qiyuan Tian, Berkin Bilgic, Qiuyun Fan, Congyu Liao, Chanon Ngamsombat, Yuxin Hu, Thomas Witzel, Kawin Setsompop, Jonathan R Polimeni, and Susie Y Huang. Deepdti: High-fidelity six-direction diffusion tensor imaging using deep learning. *NeuroImage*, 219:117017, 2020.

Qiyuan Tian, Ziyu Li, Qiuyun Fan, Jonathan R Polimeni, Berkin Bilgic, David H Salat, and Susie Y Huang. Sdndti: Self-supervised deep learning-based denoising for diffusion tensor mri. *Neuroimage*, 253:119033, 2022.

Gregory Vaksman and Michael Elad. Patch-craft self-supervised training for correlated image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5795–5804, 2023.

David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, and Kamil Ugurbil. The wu-minn human connectome project: an overview. *neuroimage*, 80:62, 2013.

Zichun Wang, Ying Fu, Ji Liu, and Yulun Zhang. Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18156–18165, 2023.

Linhai Wu, Lihui Wang, Zeyu Deng, Yuemin Zhu, and Hongjiang Wei. Replace2self: Self-supervised denoising based on voxel replacing and image mixing for diffusion mri. *IEEE Transactions on Medical Imaging*, 44(7):2878–2891, 2025. doi: 10.1109/TMI.2025.3552611.

Tiange Xiang, Mahmut Yurt, Ali B Syed, Kawin Setsompop, and Akshay Chaudhari. DDM$ˆ2$: Self-supervised diffusion MRI denoising with generative diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0vqjc50HfcC.

Yuqian Zhou, Jianbo Jiao, Haibin Huang, Yang Wang, Jue Wang, Honghui Shi, and Thomas Huang. When awgn-based denoiser meets real noises. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13074–13081, 2020.

AI USE STATEMENT

We used ChatGPT for non-scientific writing assistance, including grammar correction, sentence rephrasing, and improving clarity in the introduction and discussion sections. We also used it to help debug implementation details of the released code. No AI tool was used to generate novel ideas, experimental results, figures, or references. All AI-assisted outputs were carefully checked and edited by the authors, who retain full responsibility for the manuscript.

A Appendix

A.1 Proof of the FA-dependent Angular Lipschitz Bound

We derive the angular Lipschitz bound used in the section titled "Region-Adaptive Angular Shuffling for Reducing Structure Bias" Under the diffusion tensor model, the noise-free DW signal at voxel i along the unit gradient direction $\mathbf{g} \in \mathbb{S}^2$ is

$$S_i(\mathbf{g}) = S_0(i) \exp\left(-b\, \mathbf{g}^\mathsf{T} \mathbf{D}(i) \mathbf{g}\right), \tag{A.1}$$

where $\mathbf{D}(i)$ is the positive semidefinite diffusion tensor, $S_0(i)$ is the baseline signal, and b is the diffusion weighting. Let $\lambda_1(i)$, $\lambda_2(i)$, and $\lambda_3(i)$ be the eigenvalues of $\mathbf{D}(i)$, collected in $\boldsymbol{\lambda}(i) = [\lambda_1(i), \lambda_2(i), \lambda_3(i)]^\mathsf{T}$.

For brevity, we omit the voxel index i in the following derivation. Define

$$h(\mathbf{g}) = \mathbf{g}^\mathsf{T} \mathbf{D} \mathbf{g}. \tag{A.2}$$

For any unit tangent vector $\mathbf{u}$ satisfying $\mathbf{u}^\mathsf{T}\mathbf{g} = 0$ and $\|\mathbf{u}\|_2 = 1$, the directional derivative of $h(\mathbf{g})$ on the sphere is

$$\frac{\mathrm{d}}{\mathrm{d}t} h\big(\mathbf{g}(t)\big)\bigg|_{t=0} = 2\mathbf{u}^\mathsf{T}\mathbf{D}\mathbf{g}, \tag{A.3}$$

where $\mathbf{g}(t)$ is a unit-speed geodesic satisfying $\mathbf{g}(0) = \mathbf{g}$ and $\mathbf{g}'(0) = \mathbf{u}$. Since $\mathbf{u}^\mathsf{T}\mathbf{g} = 0$, we have

$$\begin{aligned} \left|2\mathbf{u}^\mathsf{T}\mathbf{D}\mathbf{g}\right| &= \left|2\mathbf{u}^\mathsf{T}\left(\mathbf{D} - c\mathbf{I}\right)\mathbf{g}\right| \\ &\le \lambda_{\max}(\mathbf{D}) - \lambda_{\min}(\mathbf{D}), \end{aligned} \tag{A.4}$$

where $c = [\lambda_{\max}(\mathbf{D}) + \lambda_{\min}(\mathbf{D})]/2$. Therefore, the Rayleigh quotient $h(\mathbf{g})$ is Lipschitz continuous on the sphere, with Lipschitz constant $\lambda_{\max}(\mathbf{D}) - \lambda_{\min}(\mathbf{D})$.

Using the chain rule for (A.1), the directional derivative of the DW signal is

$$\frac{\mathrm{d}}{\mathrm{d}t} S\big(\mathbf{g}(t)\big)\bigg|_{t=0} = -bS\big(\mathbf{g}\big) \frac{\mathrm{d}}{\mathrm{d}t} h\big(\mathbf{g}(t)\big)\bigg|_{t=0}. \tag{A.5}$$

Because $\mathbf{D}$ is positive semidefinite, $S(\mathbf{g}) \le S_0$. Combining this fact with (A.4) gives

$$\left|\frac{\mathrm{d}}{\mathrm{d}t} S\big(\mathbf{g}(t)\big)\bigg|_{t=0}\right| \le bS_0 \left[\lambda_{\max}(\mathbf{D}) - \lambda_{\min}(\mathbf{D})\right]. \tag{A.6}$$

For two antipodally equivalent diffusion directions $\mathbf{g}_q$ and $\mathbf{g}_r$, integrating (A.6) along their shortest geodesic yields

$$\begin{aligned} |S_i(\mathbf{g}_r) - S_i(\mathbf{g}_q)| &\le bS_0(i) \left[\lambda_{\max}(i) - \lambda_{\min}(i)\right] d_{\mathbb{S}^2}(\mathbf{g}_r, \mathbf{g}_q), \\ d_{\mathbb{S}^2}(\mathbf{g}_r, \mathbf{g}_q) &= \arccos\left(|\langle \mathbf{g}_r, \mathbf{g}_q \rangle|\right). \end{aligned} \tag{A.7}$$

The absolute inner product is used because $S_i(\mathbf{g}) = S_i(-\mathbf{g})$ under the tensor model.

It remains to relate the eigenvalue range to FA. Let

$$\bar{\lambda}(i) = \frac{\lambda_1(i) + \lambda_2(i) + \lambda_3(i)}{3}, \qquad \boldsymbol{\delta}(i) = \boldsymbol{\lambda}(i) - \bar{\lambda}(i)\mathbf{1}. \tag{A.8}$$

The fractional anisotropy is defined as

$$\mathrm{FA}(i) = \sqrt{\frac{3}{2}} \frac{\|\boldsymbol{\delta}(i)\|_2}{\|\boldsymbol{\lambda}(i)\|_2}. \tag{A.9}$$

For any pair of eigenvalues $\lambda_m(i)$ and $\lambda_n(i)$,

$$\begin{aligned} |\lambda_m(i) - \lambda_n(i)| &= |\delta_m(i) - \delta_n(i)| \\ &\le \sqrt{2}\, \|\boldsymbol{\delta}(i)\|_2 \\ &= \frac{2}{\sqrt{3}}\, \|\boldsymbol{\lambda}(i)\|_2\, \mathrm{FA}(i), \end{aligned} \tag{A.10}$$

where the second line follows from the Cauchy–Schwarz inequality. In particular, by choosing m and n as the indices of the largest and smallest eigenvalues, respectively, we obtain

$$\lambda_{\max}(i) - \lambda_{\min}(i) \le \frac{2}{\sqrt{3}}\, \|\boldsymbol{\lambda}(i)\|_2\, \mathrm{FA}(i). \tag{A.11}$$

Substituting (A.11) into (A.7) gives

$$|S_i(\mathbf{g}_r) - S_i(\mathbf{g}_q)| \le C(i)\,\mathrm{FA}(i)\,d_{\mathbb{S}^2}(\mathbf{g}_r, \mathbf{g}_q), \tag{A.12}$$

where

$$C(i) = \frac{2}{\sqrt{3}}\,bS_0(i)\|\boldsymbol{\lambda}(i)\|_2. \tag{A.13}$$

Consequently, the local angular Lipschitz constant satisfies

$$L_{\mathrm{ang}}(i) \le \frac{2}{\sqrt{3}}\,bS_0(i)\|\boldsymbol{\lambda}(i)\|_2\,\mathrm{FA}(i) = C(i)\,\mathrm{FA}(i), \tag{A.14}$$

which proves the bound used for the FA-guided angular shuffling strategy.

A.2 Detailed Network Architecture

This appendix describes the proposed dual-branch network. The network takes the tensor-fitted DW image y^a and the shuffled noisy DW image y^b as inputs and is applied in both directions. To predict $\hat{x}^b$, the blind-spot branch processes the center-excluded neighborhood Z_p^b, while the auxiliary branch extracts complementary information from y^a. Similarly, to predict $\hat{x}^a$, the blind-spot branch processes Z_p^a and the auxiliary branch processes y^b.

For target location p and prediction direction $v \in \{a, b\}$, let $\bar{v}$ denote the other view. Consistent with Eq. (12), the blind-spot and auxiliary branches first produce separate predictions, which are then combined at the output:

$$\begin{aligned} \hat{x}_p^b &= \alpha_p^b f_\theta^{\mathrm{blind}}(Z_p^b) + (1-\alpha_p^b) f_\theta^{\mathrm{aux}}(y^a)_p, \\ \hat{x}_p^a &= \alpha_p^a f_\theta^{\mathrm{blind}}(Z_p^a) + (1-\alpha_p^a) f_\theta^{\mathrm{aux}}(y^b)_p. \end{aligned} \tag{A.15}$$

where $\alpha_p^v \in [0,1]$ is the adaptive fusion weight at location p. The same network parameters are used for both prediction directions.

Blind-spot Branch For each target location p, the blind-spot branch processes the center-excluded neighborhood Z_p^v. An initial feature embedding is extracted using a 1×1 convolution:

$$F_{\mathrm{blind},p}^{v,0} = \sigma\left(W_{\mathrm{blind}}^{\mathrm{in}} * Z_p^v + b_{\mathrm{blind}}^{\mathrm{in}}\right), \tag{A.16}$$

where $*$ denotes convolution and $\sigma(\cdot)$ denotes the ReLU activation function.

The blind-spot branch excludes the center pixel from its receptive field. Let $\mathbf{p}$ denote a spatial offset in the convolution kernel support $\mathcal{K}$, where $\mathbf{0}$ represents the kernel center. The masked convolution kernel is defined as

$$\widetilde{W}(\mathbf{p}) = \begin{cases} 0, & \mathbf{p} = \mathbf{0}, \\ W(\mathbf{p}), & \mathbf{p} \ne \mathbf{0}. \end{cases} \tag{A.17}$$

Accordingly, the initial blind-spot feature at spatial location $\mathbf{r}$ is

$$F_{\mathrm{blind},p}^{v,1}(\mathbf{r}) = \sum_{\mathbf{p}\in\mathcal{K}} \widetilde{W}_{\mathrm{blind}}(\mathbf{p}) F_{\mathrm{blind},p}^{v,0}(\mathbf{r}+\mathbf{p}) + b_{\mathrm{blind}}. \tag{A.18}$$

The resulting features are further processed by a sequence of residual dilated blind-spot blocks:

$$\begin{aligned} F_{\mathrm{blind},p}^{v,l+1} = F_{\mathrm{blind},p}^{v,l} &+ W_{\mathrm{blind},2}^{l} * \sigma\left(W_{\mathrm{blind},1,d}^{l} * F_{\mathrm{blind},p}^{v,l} + b_{\mathrm{blind},1}^{l}\right) \\ &+ b_{\mathrm{blind},2}^{l}. \end{aligned} \tag{A.19}$$

where $W_{\mathrm{blind},1,d}^{l}$ denotes a dilated blind-spot convolution and $W_{\mathrm{blind},2}^{l}$ denotes a 1×1 convolution. The dilated convolutions enlarge the receptive field and enable the branch to infer the center signal from its surrounding context without directly using the observed center-pixel value. The final output feature of this branch is denoted by $F_{\mathrm{blind},p}^{v}$.

Auxiliary Branch The auxiliary branch takes the opposite view $y^{\bar{v}}$ as input. Unlike the blind-spot branch, this branch is not subject to the center-exclusion constraint, since its purpose is to provide complementary structural information from the cross-view observation. Its initial feature embedding is given by

$$F_{\mathrm{aux},p}^{\bar{v},0} = \sigma\left(W_{\mathrm{aux}}^{\mathrm{in}} * y^{\bar{v}} + b_{\mathrm{aux}}^{\mathrm{in}}\right). \tag{A.20}$$

A series of standard residual convolutional blocks is then used to extract local image content and structural features:

$$F^{\bar{v},l+1}_{\mathrm{aux},p} = F^{\bar{v},l}_{\mathrm{aux},p} + W^{l}_{\mathrm{aux},2} * \sigma\left(W^{l}_{\mathrm{aux},1} * F^{\bar{v},l}_{\mathrm{aux},p} + b^{l}_{\mathrm{aux},1}\right) + b^{l}_{\mathrm{aux},2}, \tag{A.21}$$

where $W^{l}_{\mathrm{aux},1}$ and $W^{l}_{\mathrm{aux},2}$ are 3×3 and 1×1 convolution kernels, respectively. The output of the auxiliary branch is denoted by $F^{\bar{v}}_{\mathrm{aux},p}$.

For example, when predicting $\hat{x}^b$, the blind-spot branch processes y^b and the auxiliary branch processes y^a. Conversely, when predicting $\hat{x}^a$, the blind-spot branch processes y^a and the auxiliary branch processes y^b.

Adaptive Fusion Network The adaptive fusion network estimates α^v_p from the blind-spot feature $F^{v}_{\mathrm{blind},p}$ and the cross-view auxiliary feature $F^{\bar{v}}_{\mathrm{aux},p}$. The two feature maps are concatenated along the channel dimension, and the gating map is computed as

$$\alpha^v_p = \mathrm{Sigmoid}\left(W_{g2} * \sigma\left(W_{g1} * [F^{v}_{\mathrm{blind},p}, F^{\bar{v}}_{\mathrm{aux},p}] + b_{g1}\right) + b_{g2}\right), \tag{A.22}$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation. The gate is computed for each target location and controls the relative contribution of the two branch predictions. Each branch feature is mapped to an image prediction by the shared reconstruction tail h_θ, composed of 1×1 convolutions:

$$h_\theta(F) = W_{t3} * \sigma\left(W_{t2} * \sigma\left(W_{t1} * F + b_{t1}\right) + b_{t2}\right) + b_{t3}. \tag{A.23}$$

The branch predictions are then fused at the output:

$$\hat{x}^v_p = \alpha^v_p h_\theta(F^{v}_{\mathrm{blind},p})_p + (1 - \alpha^v_p) h_\theta(F^{\bar{v}}_{\mathrm{aux},p})_p. \tag{A.24}$$

No global residual connection is used within either branch's reconstruction tail.

A.3 Detailed dataset description

Datasets Three datasets were used in this study: one simulated dataset and two real in vivo datasets. For the simulated experiments, clean DW images were generated using the virtual fiber phantom provided by the Phantomas ISBI challenge (Caruyer et al., 2014). The simulated data were acquired with a b-value of $1000\ \mathrm{s/mm^2}$ and 64 diffusion gradient directions. The image size was $55 \times 55 \times 55$, with an isotropic spatial resolution of $1.5 \times 1.5 \times 1.5\ \mathrm{mm^3}$. To evaluate denoising performance under controlled noise conditions, Rician noise and noncentral Chi (NC-Chi) noise were separately added to the clean DW images. The noise level was set to 6%, 8% and 10% of the maximum signal intensity, and the number of coils for NC-Chi noise simulation was set to 4. The simulation equations are given in the following paragraph.

The first real dataset was obtained from the Human Connectome Project (HCP) (Van Essen et al., 2013). The DW images were acquired at three b-values: 1000, 2000, and $3000\ \mathrm{s/mm^2}$. For each b-value, 90 diffusion gradient directions were included. The image size was $145 \times 174 \times 145$, with a spatial resolution of $1.25 \times 1.25 \times 1.25\ \mathrm{mm^3}$. The second real dataset was the Sherbrooke 3-Shell dataset (Garyfallidis et al., 2014). The DW images were acquired at three b-values: 1000, 2000, and $3500\ \mathrm{s/mm^2}$, with 64 diffusion gradient directions for each b-value. The image size was $128 \times 128 \times 60$.

Noisy image simulation Noisy DW magnitudes were synthesized from clean magnitudes I using Rician and noncentral Chi (NC-Chi) models. For N receiver coils, the simulated magnitude is

$$I_n = \sqrt{(I + n^R_1)^2 + \sum_{k=2}^{N} (n^R_k)^2 + \sum_{k=1}^{N} (n^I_k)^2}, \qquad N \geq 1, \tag{A.25}$$

where n^R_k and n^I_k are independently sampled real and imaginary Gaussian components with distribution $\mathcal{N}(0, \sigma^2)$. With $N = 1$, Eq. (A.25) becomes

$$I_n = \sqrt{(I + n^R_1)^2 + (n^I_1)^2}, \tag{A.26}$$

which gives the Rician model. We use $N = 1$ for Rician and $N = 4$ for NC-Chi simulation, with σ equal to 6%, 8% and 10% of the maximum signal intensity.

A.4 DETAILS ABOUT EVALUATION METRICS

To quantitatively assess the performance of different methods on the simulation dataset, we employed peak signal-to-noise ratio(PSNR) and structural similarity index measure(SSIM).

$$\begin{aligned} \text{PSNR}(I,G) &= 20 * log_{10}\left(\frac{\text{MAX}_I}{\text{RMSE}(I,G)}\right) \\ \text{SSIM}(I,G) &= \frac{(2\mu_I\mu_j + c_1)(2\sigma_{IG} + c_2)}{(\mu_I^2 + \mu_G^2 + c_1)(\sigma_I^2 + \sigma_G^2 + c_2)} \end{aligned} \tag{A.27}$$

where MAX_I represents the maximum value of the denoised image I, RMSE the root mean square error between ground-truth image G and denoised image I, μ_I and μ_G the mean intensity of images I and G, σ_I^2 and σ_G^2 the corresponding variance, while σ_{IG} the covariance between I and G, $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ two variables determined by dynamic range L of image and two constants k_1 and k_2. Higher PSNR and SSIM indicates the better denoising performance.

Learned Perceptual Image Patch Similarity(LPIPS) measures the distance between deep feature representations extracted from the pretrained network:

$$\begin{aligned} \text{LPIPS}(I,G) = \sum_l \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \\ \left\| \mathbf{w}_l \odot \left(\hat{\phi}_l(I)_{hw} - \hat{\phi}_l(G)_{hw} \right) \right\|_2^2, \end{aligned} \tag{A.28}$$

where I denotes the denoised image produced by the proposed method, and G denotes the corresponding ground-truth clean image. $\phi_l(\cdot)$ denotes the feature map extracted from the l-th layer of the pretrained AlexNet backbone, while $\hat{\phi}_l(\cdot)$ denotes the corresponding channel-wise normalized feature representation. H_l and W_l represent the height and width of the feature map at layer l, respectively; $\mathbf{w}_l$ is a learned channel-wise weighting vector; and $\odot$ denotes element-wise multiplication. A lower LPIPS value indicates greater perceptual similarity between I and G.

Additionally, to measure the accuracy of the denoised fiber orientations, we computed the average angular error(AAE), defined as follows:

$$\text{AAE} = \frac{1}{N} \sum_{i=1}^{N} \frac{180}{\pi} \arccos\left(\left| d_i^{gt} \cdot d_i^{est} \right| \right) \tag{A.29}$$

where d_i^{gt} and d_i^{est} represents the true fiber orientation and the denoised fiber orientation at i-th voxel, $\cdot$ the dot product, and N the number of voxels in a ROI. The smaller the AAE, the more accurate the denoised fiber orientation is.

Spatial correlation coefficient Let $r(\mathbf{s})$ denote the noise value at voxel location $\mathbf{s} = (x, y, z)$ within a valid spatial region Ω.

For each spatial direction $d \in \{x, y, z\}$, adjacent voxel pairs are defined as

$$\mathcal{E}_d = \{(\mathbf{s}, \mathbf{s} + \mathbf{e}_d) : \mathbf{s} \in \Omega,\ \mathbf{s} + \mathbf{e}_d \in \Omega\}, \tag{A.30}$$

where $\mathbf{e}_d$ is a unit displacement along direction d.

For all valid neighboring pairs $\{(\mathbf{s}_i, \mathbf{t}_i)\}_{i=1}^{N_d} \in \mathcal{E}_d$, define

$$a_i = r(\mathbf{s}_i), \qquad b_i = r(\mathbf{t}_i). \tag{A.31}$$

The spatial correlation in direction d is computed using Pearson's correlation coefficient:

$$\rho_d = \frac{\sum_{i=1}^{N_d} (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{N_d} (a_i - \bar{a})^2} \sqrt{\sum_{i=1}^{N_d} (b_i - \bar{b})^2}}, \qquad d \in \{x, y, z\}, \tag{A.32}$$

where

$$\bar{a} = \frac{1}{N_d} \sum_{i=1}^{N_d} a_i, \qquad \bar{b} = \frac{1}{N_d} \sum_{i=1}^{N_d} b_i. \tag{A.33}$$

The mean spatial correlation over the three directions is

$$\rho_{\text{mean}} = \frac{1}{3} (\rho_x + \rho_y + \rho_z). \tag{A.34}$$

A value of ρ_d close to zero indicates weak spatial dependence between adjacent noise voxels, while $\rho_d > 0$ indicates positive spatial correlation.

Structure-distortion metric. Let $C(\mathbf{s})$ and $P(\mathbf{s})$ denote the intensities of the clean and processed images at voxel $\mathbf{s}$, respectively, and let $\mathcal{N}(\mathbf{s})$ denote its local spatial neighborhood. We quantify structural distortion from three complementary aspects: local intensity relations, gradient orientation, and neighborhood ordering.

For each neighbor $\mathbf{q} \in \mathcal{N}(\mathbf{s})$, we first compute the relative intensity differences

$$d_C(\mathbf{s}, \mathbf{q}) = C(\mathbf{q}) - C(\mathbf{s}), \qquad d_P(\mathbf{s}, \mathbf{q}) = P(\mathbf{q}) - P(\mathbf{s}). \tag{A.35}$$

The local structural relation preservation (LSRP) score is then defined as

$$\mathrm{LSRP}(\mathbf{s}) = \mathrm{corr}\left(\{d_C(\mathbf{s}, \mathbf{q})\}_{\mathbf{q}\in\mathcal{N}(\mathbf{s})}, \{d_P(\mathbf{s}, \mathbf{q})\}_{\mathbf{q}\in\mathcal{N}(\mathbf{s})}\right), \tag{A.36}$$

where $\mathrm{corr}(\cdot, \cdot)$ denotes the Pearson correlation coefficient. This metric evaluates whether the processed image preserves the local intensity relationships of the clean image.

We further measure local gradient-direction consistency using cosine similarity:

$$G(\mathbf{s}) = \frac{\nabla C(\mathbf{s})^{\mathsf{T}} \nabla P(\mathbf{s})}{\|\nabla C(\mathbf{s})\|_2 \|\nabla P(\mathbf{s})\|_2 + \epsilon}, \tag{A.37}$$

where $\nabla C(\mathbf{s})$ and $\nabla P(\mathbf{s})$ are the spatial gradient vectors, and $\epsilon > 0$ is a small constant ensuring numerical stability. A high value indicates that local edge orientations are preserved after processing.

Finally, let $R_C(\mathbf{q})$ and $R_P(\mathbf{q})$ denote the ranks of the clean and processed intensities, respectively, within $\mathcal{N}(\mathbf{s})$. Neighborhood rank preservation (NRP) is defined as

$$\mathrm{NRP}(\mathbf{s}) = \mathrm{corr}\left(\{R_C(\mathbf{q})\}_{\mathbf{q}\in\mathcal{N}(\mathbf{s})}, \{R_P(\mathbf{q})\}_{\mathbf{q}\in\mathcal{N}(\mathbf{s})}\right), \tag{A.38}$$

which corresponds to Spearman's rank correlation coefficient and measures the preservation of local intensity ordering.

For each metric, we report its mean over the set of valid voxels Ω:

$$\overline{M} = \frac{1}{|\Omega|} \sum_{\mathbf{s}\in\Omega} M(\mathbf{s}), \qquad M \in \{\mathrm{LSRP}, G, \mathrm{NRP}\}. \tag{A.39}$$

Higher values indicate better structural preservation, whereas lower values indicate greater distortion.

Noise Expectation Bias Let $X(\mathbf{s})$ and $Y(\mathbf{s})$ denote the clean and noisy image intensities at voxel $\mathbf{s}$, respectively. The noise realization is defined as

$$N(\mathbf{s}) = Y(\mathbf{s}) - X(\mathbf{s}). \tag{A.40}$$

To quantify whether the noise at a target voxel can be predicted from its visible surroundings, we construct a blind-spot context vector from the neighboring noisy voxels within the same slice:

$$\mathbf{Z}(\mathbf{s}) = \{Y(\mathbf{q}) \,|\, \mathbf{q} \in \mathcal{N}(\mathbf{s}),\ \mathbf{q} \neq \mathbf{s}\}, \tag{A.41}$$

where $\mathcal{N}(\mathbf{s})$ denotes the local neighborhood of $\mathbf{s}$ and the center voxel is excluded.

The context-dependent noise mean is

$$m(\mathbf{Z}(\mathbf{s})) = \mathbb{E}\left[N(\mathbf{s}) \mid \mathbf{Z}(\mathbf{s})\right]. \tag{A.42}$$

Because this conditional expectation is unknown, we approximate it using an out-of-fold ridge-regression predictor:

$$\widehat{m}(\mathbf{Z}(\mathbf{s})) \approx \mathbb{E}\left[N(\mathbf{s}) \mid \mathbf{Z}(\mathbf{s})\right]. \tag{A.43}$$

Out-of-fold prediction ensures that each estimated conditional mean is obtained from a regressor that was not fitted using that voxel, thereby reducing bias from in-sample overfitting.

We define the noise expectation bias (NEB) as the root-mean-square magnitude of the predicted conditional noise mean, normalized by the root-mean-square magnitude of the observed noise:

$$\mathrm{NEB} = \frac{\left[\frac{1}{|\Omega|}\sum_{\mathbf{s}\in\Omega} \widehat{m}\big(\mathbf{Z}(\mathbf{s})\big)^2\right]^{1/2}}{\left[\frac{1}{|\Omega|}\sum_{\mathbf{s}\in\Omega} N(\mathbf{s})^2\right]^{1/2}}, \tag{A.44}$$

where Ω denotes the set of valid voxels. If the center noise is conditionally zero-mean given the visible context, then $m(\mathbf{Z}) = 0$ and NEB approaches zero. A larger NEB indicates stronger context-dependent noise and, consequently, a greater violation of the conditional zero-mean assumption underlying blind-spot self-supervision.

A.5 Baseline and Implementation Configurations

Training configuration. All experiments were implemented in PyTorch and conducted on a single NVIDIA A6000 GPU. The proposed network was trained for 200 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 32. Unless otherwise specified, the same training configuration was used across all datasets and noise settings.

Compared methods. We compared our method with five representative DW image denoising approaches: MPPCA, Patch2Self, SAN2N, Replace2Self, and UP2-MAE. MPPCA was implemented using its publicly available MATLAB toolbox, with the search radius set to 2. Patch2Self was implemented using the API provided by DIPY, with the patch radius set to 1. For SAN2N, Replace2Self, and UP2-MAE, we used their official implementations and adopted the parameter settings recommended in the corresponding papers. All methods were evaluated using the same noisy test images and evaluation protocol to ensure a fair comparison.

Signal-dependent magnitude residual. Let p denote a spatial-angular sample with clean and noisy magnitudes I_p and $I_{n,p}$. Define its total magnitude-domain residual by

$$R_p = I_{n,p} - I_p. \tag{A.45}$$

Because the magnitude operation is nonlinear, the conditional mean of this residual depends on the clean signal:

$$b_N(I_p) := \mathbb{E}[R_p \mid I_p] = \mathbb{E}[I_{n,p} \mid I_p] - I_p. \tag{A.46}$$

We can therefore write

$$R_p = b_N(I_p) + \xi_p, \qquad \mathbb{E}[\xi_p \mid I_p] = 0, \tag{A.47}$$

where ξ_p is the centered random innovation. At high SNR, expansion of Eq. (A.25) gives

$$\begin{aligned} R_p = n^R_{1,p} + \frac{1}{2I_p}\left[\sum_{k=2}^{N}(n^R_{k,p})^2 + \sum_{k=1}^{N}(n^I_{k,p})^2\right] \\ + \mathcal{O}\left(\frac{\|\mathbf{n}_p\|^3}{I_p^2}\right), \end{aligned} \tag{A.48}$$

where $\mathbf{n}_p$ collects the Gaussian components. Consequently,

$$b_N(I_p) \approx \frac{(2N-1)\sigma^2}{2I_p}. \tag{A.49}$$

For Rician noise this is $\sigma^2/(2I_p)$, while for four-coil NC-Chi noise it is $7\sigma^2/(2I_p)$.

Dependence inherited from the clean signal. Consider neighboring samples p and p' across an ensemble of locations or data instances. Under the stated simulation, the Gaussian innovations are independent across voxels. Conditional on the clean signal field, this yields

$$\mathrm{Cov}(R_p, R_{p'} \mid I_p, I_{p'}) = 0. \tag{A.50}$$

Without conditioning on that unknown field, the law of total covariance gives

$$\begin{aligned} \mathrm{Cov}(R_p, R_{p'}) &= \mathbb{E}[\mathrm{Cov}(R_p, R_{p'} \mid I_p, I_{p'})] \\ &\quad + \mathrm{Cov}(\mathbb{E}[R_p \mid I_p], \mathbb{E}[R_{p'} \mid I_{p'}]) \\ &= \mathrm{Cov}(b_N(I_p), b_N(I_{p'})). \end{aligned} \tag{A.51}$$

Neighboring clean DW signals vary together across anatomy, so their signal-dependent bias components can induce marginal spatial dependence in the total magnitude residual. At high SNR,

$$\mathrm{Cov}(R_p, R_{p'}) \approx \frac{(2N-1)^2\sigma^4}{4}\,\mathrm{Cov}(I_p^{-1}, I_{p'}^{-1}). \tag{A.52}$$

If the clean signals are locally concentrated around positive means μ_p and $\mu_{p'}$, a first-order expansion further gives

$$\mathrm{Cov}(R_p, R_{p'}) \approx \frac{(2N-1)^2\sigma^4}{4\mu_p^2\mu_{p'}^2}\,\mathrm{Cov}(I_p, I_{p'}). \tag{A.53}$$

Thus the present simulation yields conditionally independent centered innovations ξ_p and $\xi_{p'}$, while the total magnitude residuals can be marginally dependent through their signal-dependent bias. This distinction is necessary when interpreting spatial noise-correlation measurements.

A.6 Additional Results

Results on Simulated 10% Nc-Chi Noise Dataset Fig. A.1 shows the denoising results of different methods on the simulation data obtained with Nc-Chi noise distributions and a noise level of 10%. Our method produces cleaner DW images with fewer residual artifacts while better preserving fine anatomical structures and the integrity of textured regions in the zoomed-in views. The residual maps further show that our method removes noise more effectively and introduces fewer reconstruction errors around tissue boundaries and fiber regions. The quantitative results are summarized in Table A.1. Our method achieves the best performance on all three DWI metrics, with a PSNR of 26.31 ± 0.86 dB, an SSIM of 0.950 ± 0.0030, and an LPIPS of 0.011 ± 0.0066. These results indicate that our method provides more accurate image restoration and perceptually more faithful reconstruction under the challenging NC-Chi noise model. In particular, the improvement in PSNR and LPIPS over the second-best method, UP2-MAE, demonstrates the advantage of the proposed cross-view denoising strategy in suppressing noise while retaining image content.

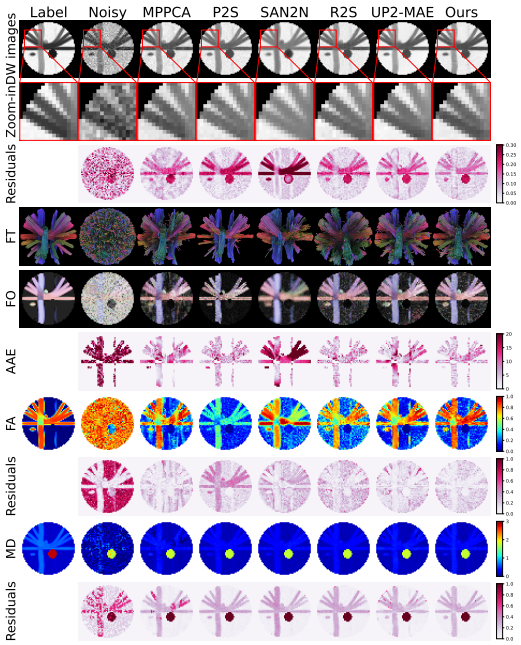


Figure A.1: Qualitative comparison of different denoising methods on the Simulation dataset under NC-Chi noise with a noise level of 10%.

The downstream DTI results provide further evidence of the effectiveness of the proposed method. Our method obtains the lowest errors in both FA and MD, with FA MAE and MD MAE values of 0.080 ± 0.094 and 0.089 ± 0.170, respectively. Compared with the second-best results of UP2-MAE, these values correspond to reductions of approximately 3.6% and 2.2%. The proposed method also achieves the lowest AAE, $4.77° \pm 3.80°$, outperforming Patch2Self, the second-best method, which obtains $6.09° \pm 3.36°$. This corresponds to an AAE reduction of approximately 21.7%, indicating that our method better preserves the diffusion orientations and produces more coherent fiber structures.

The visualizations of fiber tractography, fiber orientations, FA/MD maps, and residual errors in Fig. A.1 are consistent with these quantitative results. Several competing methods exhibit fiber loss, incomplete bundle reconstruction, or erroneous streamlines, while our method produces fiber configurations that more closely resemble the ground truth. Similarly, the FA and MD maps reconstructed by our method show fewer artifacts and better agreement with the reference maps, particularly in fiber regions and near tissue boundaries.

Method	DWI Metrics			DTI Metrics		
	PSNR ↑	SSIM ↑	LPIPS ↓	FA MAE ↓	MD MAE ↓	AAE (°) ↓
Noisy	20.22 ± 0.24	0.759 ± 0.0147	0.073 ± 0.0109	0.287 ± 0.341	0.163 ± 0.236	25.75 ± 18.35
MPPCA	25.59 ± 0.98	0.938 ± 0.0037	0.013 ± 0.0081	0.087 ± 0.106	0.102 ± 0.185	7.57 ± 5.00
SAN2N	23.97 ± 1.90	0.947 ± 0.0080	0.028 ± 0.0175	0.116 ± 0.124	0.100 ± 0.179	15.90 ± 11.79
Patch2Self	24.97 ± 1.75	0.946 ± 0.0054	0.018 ± 0.0122	0.138 ± 0.154	0.098 ± 0.175	6.09 ± 3.36
Replace2Self	25.79 ± 1.43	0.948 ± 0.0029	0.016 ± 0.0106	0.108 ± 0.124	0.098 ± 0.174	6.60 ± 4.72
UP2-MAE	26.09 ± 1.10	0.947 ± 0.0028	0.015 ± 0.0103	0.083 ± 0.105	0.091 ± 0.167	7.34 ± 5.50
Ours	$\mathbf{26.31 \pm 0.86}$	$\mathbf{0.950 \pm 0.0030}$	$\mathbf{0.011 \pm 0.0066}$	$\mathbf{0.080 \pm 0.094}$	$\mathbf{0.089 \pm 0.170}$	$\mathbf{4.77 \pm 3.80}$

Table A.1: Quantitative comparison of different methods under 10% NC-Chi noise in terms of DWI and DTI metrics. The best results are highlighted in bold.

Quantitative Results on Simulated 6% and 8% Rician/Nc-Chi Noise Dataset Fig. A.2 compares the denoising performance under Rician and NC-Chi noise at noise levels of 6% and 8%. As the noise level increases, all methods exhibit lower PSNR and SSIM, together with higher LPIPS, confirming the increased difficulty of recovering high-quality DW images from stronger noise. Our method achieves the highest or comparable PSNR across both noise distributions and noise levels.

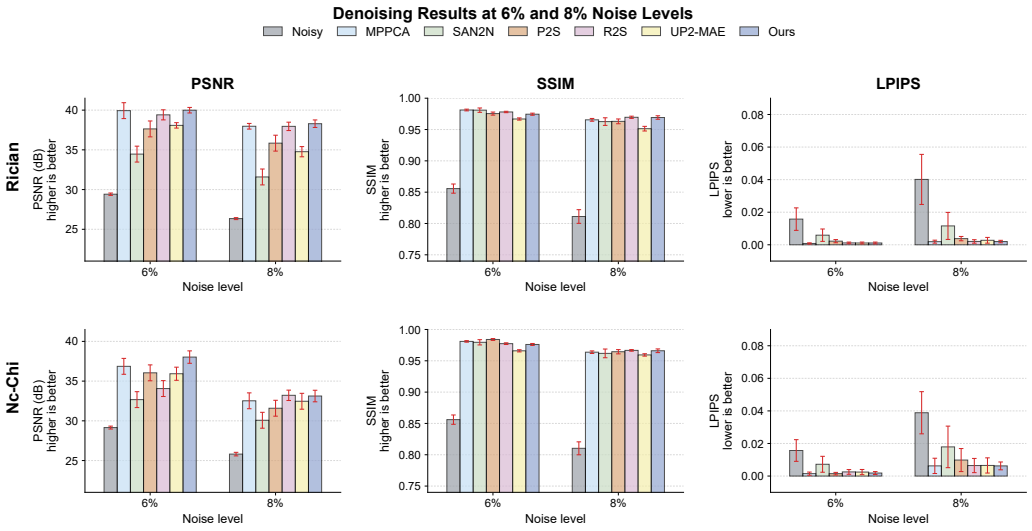


Figure A.2: Quantitative comparison of different denoising methods on the Simulation dataset under Rician and NC-Chi noise with noise level of 6% and 8%.

The improvement is particularly evident under NC-Chi noise at 6%, where our method clearly outperforms the competing approaches, and it remains among the best-performing methods when the noise level increases to 8%.

For SSIM and LPIPS, MPPCA and several learning-based methods obtain competitive results in individual settings. However, MPPCA shows lower PSNR, especially at the higher noise level, while SAN2N exhibits a pronounced performance degradation and larger variation under stronger noise. In contrast, our method consistently maintains high SSIM and low LPIPS while preserving superior PSNR. The relatively small error bars further indicate stable performance across test samples. These results demonstrate that the proposed method provides a favorable balance between signal fidelity, structural preservation, and perceptual quality, and generalizes robustly across Rician and NC-Chi noise distributions.