JET-DIFF: JOINT-ENCODING TENSOR DIFFUSION MODEL FOR ACCURATE DTI RECONSTRUCTION FROM SPARSE DWIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion Tensor Imaging (DTI) is an advanced Magnetic resonance imaging (MRI) technique for characterizing white matter microstructure. Conventional DTI protocols require multiple diffusion-weighted imaging (DWI) acquisitions across numerous directions, resulting in long scan times, motion artifacts, patient discomfort, and reduced clinical utility. Current deep learning approaches frequently yield diffusion tensors that are anatomically inconsistent or physically implausible. We introduce Joint-Encoding Tensor Diffusion (JET-Diff), a framework that synthesizes the full six-component diffusion tensor in 3D. Specifically, we propose a Multi-Tensor Latent Diffusion (MTLD) model that learns a shared latent distribution between DWIs and DTIs, enforcing both anatomical fidelity and physical plausibility. MTLD leverages a novel anatomical autoencoder to disentangle structural information from tensor properties, yielding a compact and expressive latent space optimized for generative performance. Experiments conducted on the Human Connectome Project (HCP) dataset demonstrate that JET-Diff significantly improves reconstruction accuracy and generates diffusion tensors that support more reliable downstream tractography.

1 INTRODUCTION

Diffusion Tensor Imaging (DTI) is a Magnetic Resonance Imaging (MRI) technique that quantifies anisotropic water diffusion, enabling non-invasive characterization of white matter microstructure (Basser et al., 1994; Le Bihan et al., 2001). It supports mapping of neural pathways and extraction of clinically relevant biomarkers across neurological disorders (Behrens et al., 2007; Andica et al., 2020). However, its clinical adoption is constrained by long acquisition times (Le Bihan et al., 2001). High-quality tensor estimation often requires more than thirty Diffusion-Weighted Images (DWIs) to adequately sample the diffusion signal (Mukherjee et al., 2008). This prolongs scans, which is a source of patient discomfort and a strain on clinical resources, and increases susceptibility to motion artifacts that degrade tensor accuracy (O'Donnell & Westin, 2011). Developing methods that can learn this prior to reconstruct reliable tensors from a substantially reduced number of DWIs could streamline millions of routine scans performed annually, improving both efficiency and diagnostic accuracy.

Reconstructing the six independent components of the diffusion tensor from a sparse set of DWIs is a severely ill-posed inverse problem (Lenglet et al., 2009). Traditional fitting methods like least-squares are mathematically underdetermined and fail to produce reliable results. While deep generative models have emerged as a promising data-driven solution (Tian et al., 2020; Li et al., 2021; Zhang et al., 2024), existing approaches suffer from critical limitations that compromise anatomical fidelity and physical plausibility. Many models operate on a 2D, slice-by-slice basis, disregarding volumetric continuity of neural structures and leading to anatomical inconsistencies in reconstructed 3D volumes. Others directly synthesize DTI-derived parameter maps, such as fractional anisotropy (FA) or mean diffusivity (MD), but this bypasses reconstruction of the full diffusion tensor and fails to enforce the physical constraints of diffusion imaging, since scalar maps are secondary quantities. Most critically for latent-based models, autoencoders often produce entangled latent representations, forcing a single information bottleneck to capture both fine-grained anatomical detail and complex

tensor characteristics, which induces an inherent trade-off between compression efficiency and reconstruction fidelity (Higgins et al., 2017; Chen et al.).

To overcome these limitations, we propose Joint-Encoding Tensor Diffusion (JET-Diff), a novel framework for high-fidelity, physically plausible DTI synthesis from sparse measurements. Our core contribution is a Multi-Tensor Latent Diffusion (MTLD) strategy that models the input DWI and output DTI components as a single, unified entity. By learning the joint distribution of a multi-component latent tensor, JET-Diff captures the intrinsic physical and statistical relationships between anatomy and microstructure, leading to a more robust and coherent synthesis.

Our framework is implemented as a carefully designed two-stage, fully 3D architecture. First, we introduce an Anatomical Autoencoder based on the principle of information decoupling. By providing anatomical context directly to the decoder, the latent space is freed to exclusively encode essential tensor characteristics, yielding a more efficient and expressive representation. Second, our conditional MTLD model is trained within this high-fidelity latent space to generate the complete tensor. By operating volumetrically and modeling the joint distribution from a disentangled latent space, JET-Diff produces high-resolution, physically plausible DTI volumes that remain highly consistent with the input anatomy, demonstrating substantial improvements over existing diffusion tensor reconstruction methods.

2 Related Work and Background

2.1 Diffusion Tensor Model

Diffusion Tensor Imaging (DTI) is a foundational MRI technique that quantifies the anisotropic diffusion of water molecules in biological tissues, particularly the brain's white matter. The framework, introduced by Basser et al. (1994), models the diffusion process in each voxel using a 3×3 symmetric positive semi-definite tensor, $\mathbf{D}$. This tensor linearly relates the measured diffusion-weighted signal to the applied diffusion-sensitizing gradients, as described by the Stejskal-Tanner equation (Stejskal & Tanner, 1965):

$$S(\mathbf{g}) = S_0 \exp(-b\mathbf{g}^T\mathbf{D}\mathbf{g})$$

This equation forms the physical basis for estimating the diffusion tensor from a series of diffusion-weighted measurements. Further details on the equation's parameters, tensor estimation, and derived metrics are provided in the appendix A.

2.2 DTI Reconstruction from Sparse Acquisitions

The problem of reconstructing a diffusion tensor from an insufficient number of Diffusion-Weighted Images (DWIs) is a classic, ill-posed inverse problem (Tuch, 2004). Early approaches relied on linear or weighted linear least-squares fitting, which are computationally simple but highly unstable and sensitive to noise in low-signal regimes (Basser et al., 1994). Model-based approaches leveraged compressed sensing theory to exploit sparsity priors, with Knoll et al. (2015) introducing reconstruction that applied Total Variation constraints to preserve spatial coherence.

The advent of deep learning revolutionized DTI reconstruction. SuperDTI (Li et al., 2021) demonstrated that convolutional neural networks could directly map from sparse DWIs to diffusion parameter maps, achieving remarkable reconstruction quality from as few as six gradient directions. FlexDTI (Wu et al., 2024) advanced this by introducing dynamic convolution kernels to embed gradient direction information, enabling reconstruction from flexible gradient schemes. However, most existing methods suffer from fundamental limitations: slice-wise processing ignores anatomical context, and direct synthesis of scalar maps independently can violate physical consistency, as these metrics should derive from a single underlying tensor.

2.3 Generative Diffusion Models for Medical Imaging

Denoising Diffusion Probabilistic Models (DDPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020) have emerged as state-of-the-art generative models, with significant recent applications in medical imaging. These models have proven effective for tasks such as accelerated MRI reconstruction (Chung & Ye, 2022) and high-resolution 3D volume synthesis (Wang et al., 2025).

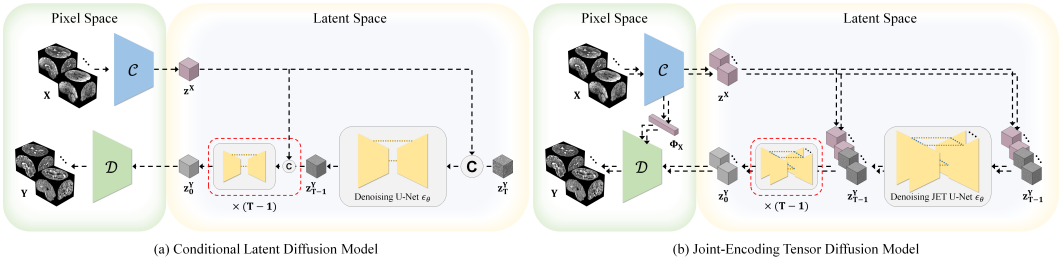


Figure 1: **Overview of the JET-Diff framework.** (a) A standard Latent Diffusion Model applies the diffusion process to DTI latents $\{\mathbf{z}_\mathbf{c}^\mathbf{Y}\}$, conditioned on DWI latents $\{\mathbf{z}_\mathbf{c}^\mathbf{X}\}$ via concatenation. (b) Our proposed JET-Diff performs joint modeling by applying the diffusion process to the combined latents of both DWIs and DTIs, $\{\mathbf{z}_\mathbf{c}^\mathbf{X}, \mathbf{z}_\mathbf{c}^\mathbf{Y}\}$.

Within the domain of diffusion MRI, recent efforts have applied generative frameworks to denoising and reconstruction. Several self-supervised methods leverage diffusion models to restore signal quality from noisy acquisitions (Xiang et al.; Wu et al.). More directly related to our task, Diff-DTI (Zhang et al., 2024) was the first to employ a diffusion model for rapid DTI reconstruction from sparse DWIs. Its approach conditions the generative process on sparse DWI features to synthesize DTI-derived scalar maps like fractional anisotropy (FA) and mean diffusivity (MD). While Diff-DTI achieves impressive results, its reliance on an explicit guidance mechanism to generate secondary parameter maps bypasses the synthesis of the fundamental diffusion tensor. In contrast, our approach learns the joint latent distribution of the input DWIs and the full six-component tensor, enabling the direct synthesis of the tensor components from which physically-consistent parameter maps are then calculated, all without the need for explicit guidance.

3 Method: Joint-Encoding Tensor Diffusion (JET-Diff)

This section details the Joint-Encoding Tensor Diffusion (JET-Diff) framework. We introduce a variant of the Latent Diffusion Model (LDM) (Rombach et al., 2022) that learns the joint distribution of sparse DWI inputs and their corresponding DTI fields, thereby promoting anatomical and physical plausibility.

3.1 Problem Definition and Overview

The primary objective is to reconstruct a complete diffusion tensor field from a minimal set of Diffusion-Weighted Images (DWIs), which is a severely ill-posed inverse problem. More specifically, the input $\mathbf{X}$ be the set of four DWI volumes, $\mathbf{X} = \{\mathbf{X_c}\}_{c=1}^4$, where each component $\mathbf{X_c} \in \mathbb{R}^{H\times W\times D}$ consists of one non-diffusion-weighted image ($b = 0$) and three DWI volumes. The desired output $\mathbf{Y}$ is the set of six diffusion tensor component volumes, $\mathbf{Y} = \{\mathbf{Y_c}\}_{c=1}^6$, where each component $\mathbf{Y_c} \in \mathbb{R}^{H\times W\times D}$ represents one of the unique tensor elements $(D_{xx}, D_{yy}, D_{zz}, D_{xy}, D_{xz}, D_{yz})$.

Our proposed framework, JET-Diff, addresses this challenge with a two-stage approach, illustrated in Figure 1. The first stage involves an Anatomical Autoencoder, composed of a tensor property encoder $\mathcal{E}$, an anatomical conditioner $\mathcal{C}$, and a DWI-aided decoder $\mathcal{D}$. This stage learns a compact latent representation of the tensor field, ensuring anatomical consistency by explicitly conditioning the synthesis process on DWI features. The second stage employs a Multi-Tensor Latent Diffusion Model (MTLD), a conditional diffusion model that generates the tensor within this latent space. Its key characteristic is the modeling of the joint distribution of the input DWIs and output DTI components.

3.2 Anatomical Autoencoder for High-Fidelity Latent Representation

The foundation of our generative framework is an autoencoder that maps high-dimensional tensor data into a compact latent space. The quality of this latent space is critical, as the performance of the subsequent diffusion model is bounded by the autoencoder's fidelity (Higgins et al., 2017;

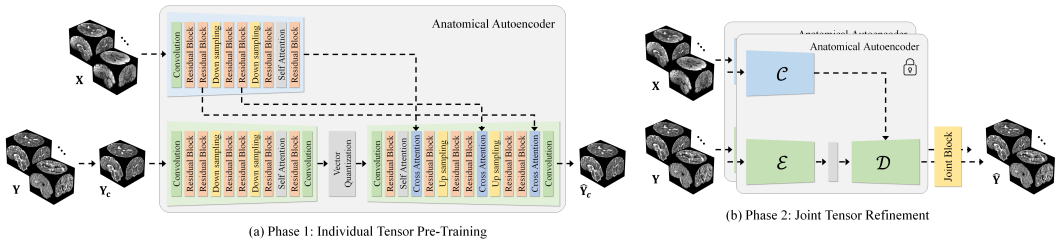


Figure 2: **Anatomical Autoencoder architecture and training.** (a) **Phase 1: Independent Pre-Training.** The encoder ($\mathcal{E}$) and conditioner ($\mathcal{C}$) respectively extract DTI latent codes and DWI anatomical features. The decoder ($\mathcal{D}$) reconstructs the DTI ($\hat{\mathbf{Y}}$) by fusing these via cross-attention. (b) **Phase 2: Joint Refinement.** With the main network frozen, a lightweight Joint Block is fine-tuned to model inter-correlations among all six tensor components.

Chen et al.). Standard autoencoders are ill-suited for this task because they force a single bottleneck to encode both the tensor's physical properties and complex anatomical structure. This entangled representation is inefficient and prone to loss of fine details. Our Anatomical Autoencoder, depicted in Figure 2, addresses this limitation through a design centered on information decoupling.

3.2.1 DWI-Aided Decoder for Information Decoupling

A key feature of our autoencoder is the principle of disentangling the latent representation of the tensor's properties (*what*) from its anatomical context (*where*). A conventional autoencoder must compress both into its latent code, creating a significant bottleneck that can lead to anatomical misalignment.

Our DWI-Aided Decoder, $\mathcal{D}$, resolves this by decoupling these responsibilities. The encoder $\mathcal{E}$ learns a highly efficient latent code $\mathbf{z}^{\mathbf{Y}}$ representing only the tensor's intrinsic properties. The anatomical context is extracted by a conditioner $\mathcal{C}$ directly from the input DWI stack $\mathbf{X}$ as a feature pyramid $\Phi_{\mathbf{X}} = \{\phi^l_{\mathbf{X}}\}_{l=1}^{L}$. During decoding, the decoder fuses the compact latent code $\mathbf{z}^{\mathbf{Y}}$ with these anatomical features $\Phi_{\mathbf{X}}$ at each resolution level. This fusion is achieved using cross attention blocks that employ a multi-axis cross attention (Tu et al., 2022) to maintain linear computational complexity, a critical requirement for processing high resolution medical images. This design allows the latent space to achieve a higher compression ratio while enabling the decoder to produce a final output $\hat{\mathbf{Y}} = \mathcal{D}(\mathbf{z}^{\mathbf{Y}}, \Phi_{\mathbf{X}})$ with superior fidelity. This high-quality autoencoder is the key to enabling our high-performance latent diffusion model.

3.2.2 Decoupled Joint Refinement for Tensor Consistency

While the DWI-aided decoder ensures high fidelity for individual tensor components, it does not explicitly enforce the physical cross-correlations required for a valid tensor field. To address this, we introduce a highly efficient decoupled joint refinement phase, illustrated in Figure 2b. After the initial training, we freeze the weights of the conditioner $\mathcal{C}$, the encoder $\mathcal{E}$, and the majority of the decoder $\mathcal{D}$. We then insert a lightweight Joint MLP block, composed of two fully connected layers, into the final layers of the decoder. This block operates concurrently on the feature maps for all tensor components just before the final output convolution, allowing it to explicitly model their inter-relationships. By fine-tuning only this joint block and the final convolution, we enforce tensor-wide consistency with minimal computational overhead. The result is a single, coherent, and physically plausible diffusion tensor.

3.3 Multi-Tensor Latent Diffusion (MTLD)

Input DWIs and corresponding DTI components are coupled manifestations of the same diffusion process. Building on this principle, we introduce the core generative component of our framework: the Multi-Tensor Latent Diffusion (MTLD) model (Figure 3). Operating within the high-fidelity latent space established earlier, the MTLD models their joint distribution by enabling direct interaction between the latent representations of all DWI and DTI components throughout the denoising process.

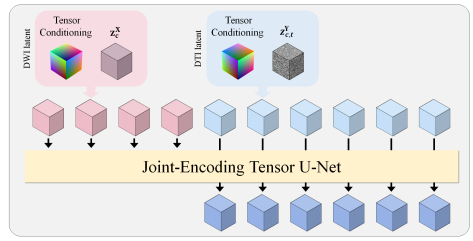
The MTLD operates on the set of latent tensors, $\{\mathbf{z}_\mathbf{c}^\mathbf{X}, \mathbf{z}_\mathbf{c}^\mathbf{Y}\}$. This set is comprised of the individual latent representations for each component volume in the input set $\mathbf{X}$ and the target set $\mathbf{Y}$, which are obtained using their respective frozen encoders. The forward diffusion process is applied independently to each latent tensor, gradually adding Gaussian noise over T timesteps (Ho et al., 2020):

$$\mathbf{z}_{\mathbf{c},t} = \sqrt{\bar{\alpha}_t}\mathbf{z}_{\mathbf{c},0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\mathbf{c}, \quad \text{for each component } c.$$

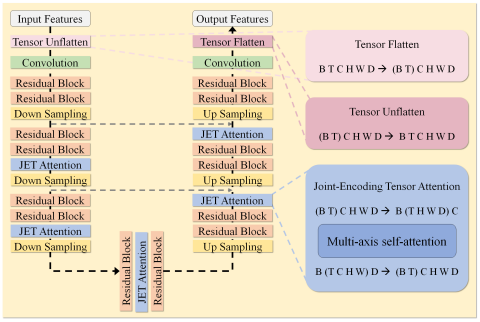
Here, $\mathbf{z}_{\mathbf{c},t}$ is the noisy version of the c-th latent tensor component, and $\bar{\alpha}_t$ denotes the predefined noise schedule. Collectively, the set of all noisy DTI latent components at timestep t is denoted as $\{\mathbf{z}_{\mathbf{c},t}^\mathbf{Y}\}$. The reverse process is learned by our proposed Joint-Encoding Tensor U-Net (JET U-Net), $\boldsymbol{\epsilon}_\theta$, which is trained to predict and remove the noise from all components simultaneously.

3.3.1 Tensor and Positional Conditioning

To enable the JET U-Net $\boldsymbol{\epsilon}_\theta$ to distinguish between the different tensor components (B_0, D_{xx}, etc.) and leverage their spatial relationships, each input latent is first augmented with explicit type and position information. As shown in Figure 3a, we employ a Tensor Conditioning module for this purpose. This module generates a conditioning representation by merging learnable tensor-specific embeddings and Fourier positional embeddings (Tancik et al., 2020). This combined representation is then concatenated along the channel dimension with its corresponding latent tensor before being fed into the JET U-Net.



(a) Multi-Tensor Latent Diffusion Model



(b) Architecture of Joint-Encoding Tensor U-Net

Figure 3: **The Multi-Tensor Latent Diffusion Model (MTLD).** (a) Both DWI ($\{\mathbf{z}_\mathbf{c}^\mathbf{X}\}$) and noisy DTI ($\{\mathbf{z}_{\mathbf{c},t}^\mathbf{Y}\}$) latents are augmented with component-type and positional information. (b) The Joint-Encoding Tensor U-Net (JET U-Net) architecture uses a JET Attention block to model interactions among all DWI and DTI latent components during denoising.

3.3.2 Unconditional Pre-training for Latent Prior

To stabilize training, we first pre-train our JET U-Net, $\boldsymbol{\epsilon}_\theta$, in an unconditional setting. In this phase, the JET Attention blocks are deactivated, and the network functions as a standard diffusion model trained on single latent tensors from the DWI and DTI sets. This step does not model cross-component correlations but instead learns a strong prior over the distribution of valid latent tensors, providing a reliable initialization for subsequent fine-tuning.

$$\mathcal{L}_{\text{pretrain}} = \mathbb{E}_{t,\{\mathbf{z}_{\mathbf{c},0}\},\boldsymbol{\epsilon}}\left[\sum_c \|\boldsymbol{\epsilon}_\mathbf{c} - \boldsymbol{\epsilon}_{\theta,\mathbf{c}}(\{\mathbf{z}_{\mathbf{c},t}\}, t)\|_2^2\right]$$

3.3.3 Conditional Fine-tuning for Guided Synthesis

After pre-training, we fine-tune the model for its primary task of generating the DTI latents ($\{\mathbf{z}_\mathbf{c}^\mathbf{Y}\}$) conditioned on the DWI latents ($\{\mathbf{z}_\mathbf{c}^\mathbf{X}\}$). In this stage, the JET (Joint-Encoding Tensor) Attention blocks are activated within the JET U-Net, $\boldsymbol{\epsilon}_\theta$. These blocks perform computationally efficient multi-axis self-attention (Tu et al., 2022), enabling the network to capture non-local correlations between DWI anatomy and DTI microstructure. The model receives the noisy latents $\{\mathbf{z}_{\mathbf{c},t}\}$ and is guided by the clean DWI latents $\{\mathbf{z}_\mathbf{c}^\mathbf{X}\}$ through the JET attention mechanism.

The objective remains to predict the noise for all channels, but now with the added guidance from the joint modeling:

$$\mathcal{L}_{\text{cond}} = \mathbb{E}_{t,\{\mathbf{z}_{\mathbf{c},0}\},\boldsymbol{\epsilon}}\left[\sum_c \|\boldsymbol{\epsilon}_\mathbf{c} - \boldsymbol{\epsilon}_{\theta,\mathbf{c}}(\{\mathbf{z}_{\mathbf{c},t}\}, t, \{\mathbf{z}_\mathbf{c}^\mathbf{X}\})\|_2^2\right]$$

By explicitly modeling the interactions between all latent variables, our MTLD robustly learns to generate DTI components that are not only statistically likely but also anatomically consistent with

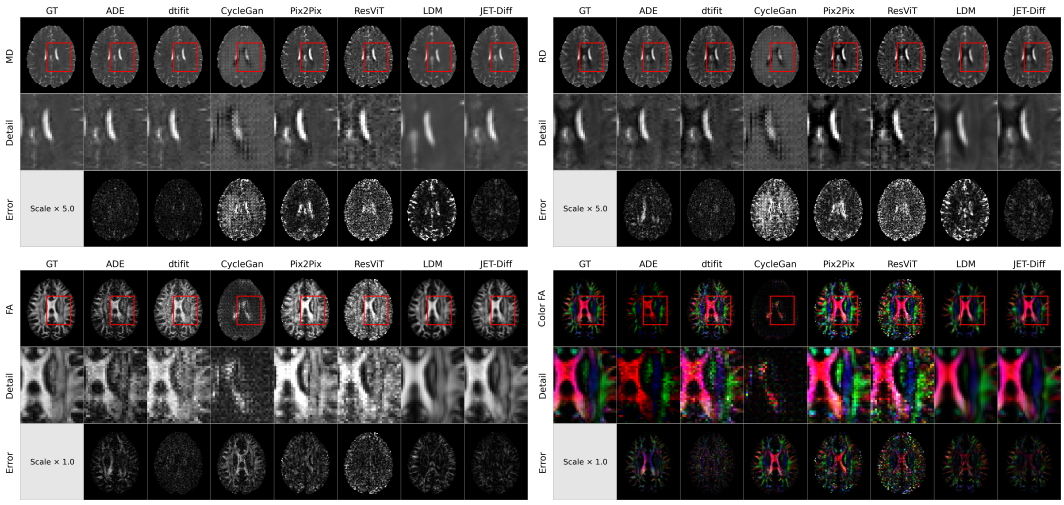


Figure 4: Qualitative comparison of DTI parameter maps (MD, RD, FA, and Color FA). For a representative subject, JET-Diff generates reconstructions with superior anatomical fidelity and lower error. Magnified insets (red box) and error maps (scaled for visibility) highlight the improved detail relative to the ground truth (GT) and competing methods.

the conditioning DWI volumes. At inference, the generated DTI latents, $\{\hat{\mathbf{z}}_c^{\mathbf{Y}}\}$, are passed to the decoder $\mathcal{D}$ to synthesize the final tensor field $\hat{\mathbf{Y}}$.

4 EXPERIMENTS

4.1 SETUPS

4.1.1 DATA AND PREPROCESSING

All experiments are conducted on diffusion MRI data from the Human Connectome Project (HCP) Young Adult dataset (Van Essen et al., 2013), utilizing DWI volumes acquired at a b-value of 1000 s/mm^2 and preprocessed with the standard HCP pipelines. Ground-truth diffusion tensors are computed for each subject via a linear least-squares fit on the full set of 90 DWI directions. All DWI volumes are resampled to 2mm isotropic resolution. The input to our model is a sparse 4-volume stack: one non-diffusion-weighted (b=0) image and the three DWI volumes whose gradient vectors are most closely aligned with the principal x, y, and z axes. The output is the complete 6-component diffusion tensor field. The full dataset of 973 subjects is partitioned into training (681), validation (97), and test (195) sets. Further details on data preparation are available in Appendix B.

4.1.2 IMPLEMENTATION DETAILS

All experiments were implemented in PyTorch (Paszke et al., 2019) and conducted on a single NVIDIA A6000 GPU. Training followed a four-stage pipeline designed to first establish a high-fidelity latent space and then train the generative model within it. The first two stages focus on the Anatomical Autoencoder: (1) independent pre-training of each tensor component and (2) a decoupled joint refinement to enforce inter-component correlations. The subsequent two stages train the Multi-Tensor Latent Diffusion model: (3) unconditional pre-training to learn a prior over the latent manifold and (4) conditional fine-tuning for guided synthesis from sparse DWI latents. Detailed architectures, loss functions, and stage-specific objectives are provided in Appendix C.

4.1.3 COMPETING METHODS

We benchmark JET-Diff against five methods: analytic diagonal estimation (ADE), a non-learning baseline that assumes a diagonal diffusion tensor by setting off-diagonal elements to zero, and four deep learning baselines: CycleGAN (Zhu et al., 2017), Pix2Pix (Isola et al., 2017), ResViT (Dalmaz

Table 1: Quantitative comparison of DTI parameter map reconstruction. Mean NMSE, PSNR, and SSIM across the test set. JET-Diff consistently outperforms other methods, particularly on FA and Color FA metrics, which are crucial for white matter analysis.

	MD			RD			FA			Color FA		
Model	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM
ADE	0.086	**26.15**	**0.969**	0.096	**27.45**	**0.970**	0.295	17.23	0.787	0.828	21.43	0.681
CycleGAN	0.142	19.76	0.781	0.182	20.23	0.770	0.590	14.23	0.576	1.237	19.69	0.594
Pix2Pix	0.087	21.90	0.930	0.103	22.71	0.931	0.425	15.72	0.790	1.372	19.29	0.694
ResViT	0.101	21.26	0.885	0.119	22.08	0.888	0.643	13.89	0.701	1.832	18.03	0.622
LDM	0.109	20.90	0.836	0.132	21.64	0.842	0.322	16.86	0.689	0.710	22.10	0.707
JET-Diff	**0.033**	26.08	0.956	**0.043**	26.58	0.952	**0.192**	**19.12**	**0.828**	**0.618**	**22.70**	**0.763**

Table 2: Quantitative comparison of diffusion tensor components. Mean PSNR and SSIM for the six independent tensor components (D_{ij}) across the test set. JET-Diff achieves the most accurate and balanced reconstruction across all components.

	D_{xx}		D_{yy}		D_{zz}		D_{xy}		D_{xz}		D_{yz}	
Model	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ADE	29.59	**0.961**	29.39	**0.960**	29.57	**0.960**	24.82	0.654	24.36	0.653	24.49	0.648
CycleGAN	25.14	0.798	24.88	0.797	25.11	0.800	21.92	0.588	21.24	0.584	21.76	0.598
Pix2Pix	27.93	0.934	27.53	0.933	27.72	0.933	24.18	0.715	24.13	0.717	26.06	0.794
ResViT	26.85	0.895	26.68	0.897	26.77	0.895	23.17	0.657	23.06	0.666	24.21	0.713
LDM	26.95	0.867	26.72	0.864	26.85	0.865	**27.56**	0.795	**27.22**	0.789	27.17	0.785
JET-Diff	**31.04**	0.953	**30.85**	0.952	**30.96**	0.953	27.46	**0.796**	27.14	**0.792**	**27.32**	**0.796**

et al., 2022), and a vanilla conditional LDM (Rombach et al., 2022). To ensure a fair comparison, all learning-based baselines are implemented with 3D networks and trained volumetrically on the same data splits and with identical input. Full descriptions are available in Appendix D.

4.2 MAIN RESULTS

4.2.1 QUALITATIVE RESULTS

Figure 4 presents a qualitative comparison of the DTI parameter maps (MD, RD, FA, and Color FA) generated by JET-Diff and competing methods for a representative subject. Each row includes whole-slice views, magnified insets, and error maps relative to the ground truth. The classical approach (ADE) introduces substantial noise and structural distortions. CycleGAN fails to restore the image entirely, while Pix2Pix and ResViT produce very noisy reconstructions with poor anatomical fidelity. The standard LDM suppresses noise more effectively but oversmooths fine structures, erasing critical white matter details. In contrast, JET-Diff achieves reconstructions that closely resemble the ground truth, effectively suppressing noise while maintaining sharp, coherent anatomy. The error maps confirm this fidelity, highlighting JET-Diff's ability to recover high-quality DTI parameters from undersampled data. Figure 7 shows the six tensor components. Competing methods exhibit noise and blurring, especially in the off-diagonal terms (D_{xy}, D_{xz}, D_{yz}), which are critical yet difficult to estimate. JET-Diff yields sharper and more coherent reconstructions across all tensor elements, providing the basis for more reliable parameter maps.

4.2.2 QUANTITATIVE RESULTS

We quantitatively evaluated all methods using NMSE, PSNR, and SSIM (Wang et al., 2004), with results summarized in Tables 1 and 2. JET-Diff achieves the strongest performance across derived parameter maps, particularly for Fractional Anisotropy (FA) and Color FA, which are highly sensitive to tensor orientation and microstructural detail.

The ADE baseline highlights important limitations of conventional metrics. It achieves deceptively high PSNR and SSIM scores for MD, RD, and the diagonal tensor elements, but only because it ignores the off-diagonal components. By fitting the smoother diagonal terms that dominate mean intensity, ADE secures favorable scores yet produces a degenerate tensor solution. This failure is reflected in its poor FA accuracy and inability to capture anisotropy, showing how conventional metrics can mask fundamental errors. The vanilla LDM baseline illustrates a different limitation.

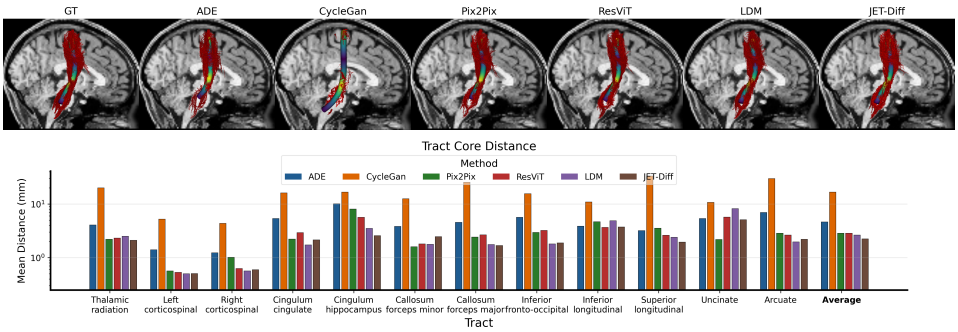


Figure 5: Tractography comparison. (Top) 3D visualization of the right corticospinal tract (CST) shows that tracts from JET-Diff tensors most closely match the ground truth. (Bottom) The mean tract core distance (mm, log scale) across 12 major white matter bundles confirms that JET-Diff yields the most geometrically accurate fiber tracking among competing methods.

It attains slightly higher PSNR on some off-diagonal elements than JET-Diff, but this reflects local voxel-wise fits rather than tensor-level coherence. JET-Diff, in contrast, achieves high accuracy on the dominant diagonal components while remaining competitive on the off-diagonals. Its joint-encoding mechanism produces a balanced reconstruction across all tensor elements, yielding more reliable parameter maps such as FA and Color FA that better reflect the underlying white matter structure.

4.2.3 Tractography comparisons

To evaluate the practical utility of the reconstructed tensors, we performed whole-brain probabilistic tractography (Garyfallidis et al., 2014; Girard et al., 2014). This task provides a stringent validation, as it depends on the coherence of all six tensor components and is highly sensitive to errors in their orientation fields. Figure 5 shows that fiber bundles generated from JET-Diff closely follow the ground truth, outperforming all competing methods both qualitatively and quantitatively. Quantitatively, JET-Diff achieves the lowest tract core distance across major white matter bundles, confirming its ability to produce tensors that reliably guide fiber tracking. Notably, while vanilla LDM achieves slightly higher PSNR on certain off-diagonal components, it performs worse in tractography, highlighting the limitation of voxel-wise metrics. ADE fails completely in this task despite favorable scores on simpler metrics, underscoring the need to assess DTI reconstruction with downstream analyses that reflect functional anatomical utility.

4.3 Ablation studies

To validate our key architectural design choices, we conducted ablation studies focused on the foundational components of our framework: the autoencoder design and the diffusion model's pre-training strategy.

4.3.1 Contribution of the DWI-Aided Decoder

To evaluate the contribution of conditioning on DWI features, we compared our Anatomical Autoencoder against a Standard Autoencoder baseline that reconstructs tensor components solely from the latent code. As shown in Table 3, the baseline performs markedly worse across all metrics and parameter maps. This confirms that forcing the latent code to represent both anatomical structure and diffusion content leads to representational entanglement and weak reconstructions. In contrast, by decoupling anatomy from the latent space and injecting multi-scale DWI features through the decoder, our architecture learns a more expressive and efficient representation, yielding consistently superior reconstructions.

4.3.2 Impact of the Joint Decoder for Tensor Consistency

We investigated the role of our decoupled refinement phase by comparing the full Anatomical Autoencoder to an ablated version without this step. As shown in Table 3, the differences in map-level

Table 3: Ablation study of the autoencoder architecture. Performance on key DTI parameter maps for the full Anatomical Autoencoder, the model without joint refinement, and a standard autoencoder baseline that is not conditioned on DWI volumes.

	MD			RD			FA			Color FA		
Model	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM
Anatomical Autoencoder (Ours)	**0.0067**	**33.09**	**0.988**	**0.0079**	**33.89**	**0.989**	**0.0931**	**22.27**	**0.904**	**0.3797**	**24.82**	**0.860**
Anatomical Autoencoder (w/o Joint Refinement)	0.0069	32.96	**0.988**	0.0081	33.77	**0.989**	0.0939	22.23	0.903	0.3826	24.79	0.859
Standard Autoencoder (Baseline)	0.0260	27.11	0.950	0.0316	27.83	0.955	0.2005	18.93	0.792	0.5595	23.14	0.780

metrics are modest, but our full model achieves a slight and consistent improvement across all parameters. This aligns with the design of the refinement block: its purpose is not to markedly boost voxel-wise accuracy, but to model the cross-channel correlations that stabilize the tensor as a whole. Enforcing these correlations reduces the frequency of invalid tensors and ensures a more coherent tensor field, benefits that are most evident in downstream applications such as tractography.

4.3.3 Impact of Unconditional Latent Diffusion Pre-Training

Our MTLD framework begins with an unconditional pre-training stage to initialize the network on the joint DTI–DWI manifold before conditional fine-tuning. To assess its impact, we trained an ablated model entirely from scratch, skipping this initialization. As shown in Figure 6, the absence of pre-training leads to noisier and less coherent reconstructions. The magnified insets reveal disrupted and less smooth V1 orientation fields, indicating weaker anatomical consistency. These results confirm that pre-training equips the model with a stable latent representation, providing a critical foundation for the conditional stage and yielding higher-fidelity tensor reconstructions.

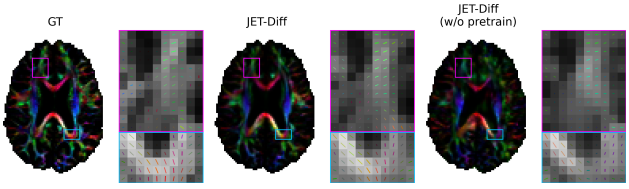


Figure 6: Ablation study on unconditional pre-training. Comparison of the full JET-Diff model against a version trained without the pre-training stage. The absence of pre-training results in noisier Color FA maps and less coherent principal eigenvector (V1) fields, as highlighted in the magnified insets.

5 Conclusion

In this work, we introduced JET-Diff, a latent diffusion framework for reconstructing high-quality diffusion tensors from critically undersampled DWI data. JET-Diff addresses key limitations of existing methods by improving anatomical coherence across volumes and capturing the correlations required for a valid tensor field. The core contribution is the Multi-Tensor Latent Diffusion (MTLD) strategy, which models the joint distribution of DWI and DTI latents within a unified generative process. This capability is enabled by the DWI-aided Anatomical Autoencoder, which separates anatomical context from tensor properties to form an efficient latent space. Extensive evaluations spanning tensor components, derived DTI parameters, and downstream tractography demonstrate that JET-Diff consistently improves reconstruction fidelity over competing approaches.

Reproducibility Statement

Our proposed method is fully reproducible. For methodology and implementation details, readers are referred to our source code, which is available in the Supplementary Material.

References

Christina Andica, Koji Kamagata, Taku Hatano, Yuya Saito, Kotaro Ogaki, Nobutaka Hattori, and Shigeki Aoki. Mr biomarkers of degenerative brain disorders derived from diffusion imaging.

Journal of Magnetic Resonance Imaging, 52(6):1620–1636, 2020.

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

Timothy EJ Behrens, H Johansen Berg, Saad Jbabdi, Matthew FS Rushworth, and Mark W Woolrich. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *neuroimage*, 34(1):144–155, 2007.

Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *The Thirteenth International Conference on Learning Representations*.

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

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022.

Eleftherios Garyfallidis, Matthew Brett, Bagrat Amirbekian, Ariel Rokem, Stefan Van Der Walt, Maxime Descoteaux, Ian Nimmo-Smith, and Dipy Contributors. Dipy, a library for the analysis of diffusion mri data. *Frontiers in neuroinformatics*, 8:8, 2014.

Gabriel Girard, Kevin Whittingstall, Rachid Deriche, and Maxime Descoteaux. Towards quantitative connectivity analysis: reducing tractography biases. *Neuroimage*, 98:266–278, 2014.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

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

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

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

Derek K Jones, Thomas R Knösche, and Robert Turner. White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion mri. *Neuroimage*, 73:239–254, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Florian Knoll, José G Raya, Rafael O Halloran, Steven Baete, Eric Sigmund, Roland Bammer, Tobias Block, Ricardo Otazo, and Daniel K Sodickson. A model-based reconstruction for undersampled radial spin-echo dti with variational penalties on the diffusion tensor. *NMR in Biomedicine*, 28(3):353–366, 2015.

Denis Le Bihan, Jean-François Mangin, Cyril Poupon, Chris A Clark, Sabina Pappata, Nicolas Molko, and Hughes Chabriat. Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 13(4):534–546, 2001.

Christophe Lenglet, Jennifer SW Campbell, Maxime Descoteaux, Gloria Haro, Peter Savadjiev, Demian Wassermann, Alfred Anwander, Rachid Deriche, G Bruce Pike, Guillermo Sapiro, et al. Mathematical methods for diffusion mri processing. *Neuroimage*, 45(1):S111–S122, 2009.

Hongyu Li, Zifei Liang, Chaoyi Zhang, Ruiying Liu, Jing Li, Weihong Zhang, Dong Liang, Bowen Shen, Xiaoliang Zhang, Yulin Ge, et al. Superdti: Ultrafast dti and fiber tractography with deep learning. *Magnetic resonance in medicine*, 86(6):3334–3347, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Partha Mukherjee, SW Chung, JI Berman, CP Hess, and RG Henry. Diffusion tensor mr imaging and fiber tractography: technical considerations. *American Journal of Neuroradiology*, 29(5): 843–852, 2008.

Lauren J O'Donnell and Carl-Fredrik Westin. An introduction to diffusion tensor image analysis. *Neurosurgery Clinics of North America*, 22(2):185, 2011.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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

Edward O Stejskal and John E Tanner. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *The journal of chemical physics*, 42(1):288–292, 1965.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

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.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pp. 459–479. Springer, 2022.

David S Tuch. Q-ball imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 52(6):1358–1372, 2004.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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

Haoshen Wang, Zhentao Liu, Kaicong Sun, Xiaodong Wang, Dinggang Shen, and Zhiming Cui. 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. *IEEE Transactions on Medical Imaging*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chenxu Wu, Qingpeng Kong, Zihang Jiang, and S Kevin Zhou. Self-supervised diffusion mri denoising via iterative and stable refinement. In *The Thirteenth International Conference on Learning Representations*.

Zejun Wu, Jiechao Wang, Zunquan Chen, Qinqin Yang, Zhen Xing, Dairong Cao, Jianfeng Bao, Taishan Kang, Jianzhong Lin, Shuhui Cai, et al. Flexdti: flexible diffusion gradient encoding scheme-based highly efficient diffusion tensor imaging using deep learning. *Physics in Medicine & Biology*, 69(11):115012, 2024.

Tiange Xiang, Mahmut Yurt, Ali B Syed, Kawin Setsompop, and Akshay Chaudhari. Ddm 2: Self-supervised diffusion mri denoising with generative diffusion models. In *The Eleventh International Conference on Learning Representations*.

Lang Zhang, Jinling He, Wang Li, Dong Liang, and Yanjie Zhu. Diff-dti: Fast diffusion tensor imaging using a feature-enhanced joint diffusion model. *IEEE Journal of Biomedical and Health Informatics*, 2024.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A Diffusion Tensor Model Details

A.1 The Stejskal-Tanner Equation Explained

The Stejskal-Tanner equation provides the foundational model for DTI (Stejskal & Tanner, 1965). The terms are defined as follows:

- S_0: The signal intensity measured in a non-diffusion-weighted acquisition (a B_0 image), where the diffusion-sensitizing gradients are turned off.
- $S(\mathbf{g})$: The signal intensity measured when a diffusion-sensitizing magnetic field gradient is applied along the direction of the unit vector $\mathbf{g}$.
- **b-value**: A scalar value that encapsulates the strength and duration of the diffusion gradients. A higher b-value results in greater signal attenuation for diffusing water molecules.

A.2 Tensor Estimation and Clinical Context

To solve for the six unknown components of the symmetric tensor $\mathbf{D}$, the Stejskal-Tanner equation must be sampled with at least six non-collinear gradient directions ($\mathbf{g}$). In clinical and research practice, many more directions (often 30 to 90 or more) are acquired to improve the accuracy and robustness of the tensor fit, especially in noisy data (Jones et al., 2013). This requirement leads to the primary clinical challenge of DTI: long acquisition times, which increase patient discomfort and sensitivity to motion artifacts.

A.3 Tensor-Derived Metrics

The diffusion tensor $\mathbf{D}$ is rarely interpreted directly. Instead, it is diagonalized to yield three eigenvalues ($\lambda_1, \lambda_2, \lambda_3$) and their corresponding eigenvectors. These represent the magnitude of diffusion in three orthogonal directions and the orientation of those directions, respectively. From these, crucial microstructural metrics are calculated (Basser et al., 1994):

- **Mean Diffusivity (MD)**: The average of the eigenvalues, $MD = (\lambda_1 + \lambda_2 + \lambda_3)/3$. It measures the overall magnitude of water diffusion in a voxel.
- **Fractional Anisotropy (FA)**: A normalized measure of the variance of the eigenvalues, indicating the degree to which diffusion is directional. An FA of 0 implies isotropic diffusion, while an FA close to 1 implies diffusion is restricted to a single direction.

These metrics are essential for the quantitative analysis of white matter integrity.

B IMAGE PROCESSING DETAILS

Ground-Truth Tensor Generation: The ground-truth DTI metrics for each subject were derived from the complete diffusion dataset, which included 18 b=0 volumes and 90 DWI volumes at b=1000 s/mm^2. Diffusion tensor fitting was performed using an ordinary linear least-squares method via the `dtifit` function in FSL (Jenkinson et al., 2012), incorporating the provided gradient nonlinearity correction files. This process yielded the full diffusion tensor, from which all ground-truth metrics, including fractional anisotropy (FA) and mean diffusivity (MD), were calculated.

Undersampled Input Selection: The 4-volume sparse input for our model was created by selecting a specific subset of DWIs. For the b=1000 s/mm^2 shell, we identified the three diffusion gradient vectors most closely aligned with the standard Cartesian axes ([1, 0, 0], [0, 1, 0], and [0, 0, 1]) by minimizing the Euclidean distance. The corresponding DWI volumes were extracted, and a single b=0 s/mm^2 volume was prepended to form the final 4-volume input stack, $\mathcal{B}$.

C JET-DIFF: IMPLEMENTATION DETAILS

Our proposed method, JET-Diff, is trained in a four-stage process designed to sequentially build the model's capabilities. The first two stages establish the high-fidelity latent space via the Anatomical Autoencoder, while the final two stages train the Multi-Tensor Latent Diffusion model to operate within that space. We use the Adam optimizer (Kingma & Ba, 2014) for the autoencoder stages and AdamW (Loshchilov & Hutter, 2017) for the diffusion model stages.

Stage 1: Anatomical Autoencoder Pre-training. We first train our autoencoder to learn a high-fidelity latent representation for each of the six tensor components independently. The encoder compresses each component into a latent space with 6 embedding dimensions and a codebook of 1024 entries. The architecture uses a base of 64 channels, channel multipliers of (1, 2, 4), and contains two residual blocks per resolution level. This stage is trained using the Adam optimizer with a learning rate of 1.0×10^{-5} and an effective batch size of 8. The objective function consists of a pixel-wise reconstruction loss and a vector-quantization commitment loss (Van Den Oord et al., 2017).

Stage 2: Decoupled Joint Refinement. After pre-training, we freeze the autoencoder weights and fine-tune a new joint decoder to enforce consistency across all six tensor components. All weights are frozen except for a new joint fusion block and the final output convolution layers of the decoder. Optimization uses the Adam optimizer with a learning rate of 1.0×10^{-4} and an effective batch size of 16. The training objective is a pixel-wise reconstruction loss combined with an adversarial loss to ensure the physical plausibility of the full tensor field.

Stage 3: Unconditional Latent Diffusion Pre-training. To provide a strong initialization for the generative model, we first pre-train the JET U-Net to model the joint distribution of the entire 10-channel latent space ($\mathbf{z} = [\mathbf{z}_B, \mathbf{z}_{\text{tensor}}]$) in an unconditional setting. The U-Net backbone has 256 base channels, channel multipliers of (1, 2, 4), two residual blocks per scale, and self-attention at multiple resolutions. The diffusion process uses a linear beta schedule (Ho et al., 2020) over 1000 steps. In this phase, the model is trained to denoise all 10 channels simultaneously, learning a robust prior over the latent manifold. The model is trained with the AdamW optimizer with a learning rate of 1.0×10^{-6} and a batch size of 8.

Stage 4: Conditional Latent Diffusion Fine-tuning. The model is then fine-tuned for the primary conditional synthesis task, initialized from the checkpoint of the unconditional pre-training phase. The forward diffusion process continues to apply Gaussian noise to the entire 10-channel latent tensor. The JET U-Net is now conditioned on the clean DWI latents ($\mathbf{z}_B$), which are concatenated to the noisy latent tensor ($\mathbf{z}_t$) as input. The model is trained to predict the noise for all 10 channels, guided by the clean DWI condition. This stage is trained using the AdamW optimizer with a learning rate of 1.0×10^{-6} and an effective batch size of 8.

D COMPETING METHODS: IMPLEMENTATION DETAILS

Unless otherwise noted, all learnable baselines are implemented as 3D networks and trained on full 3D volumes. All methods use the same data splits, input/output formats, and evaluation protocols for a fair comparison.

- **ADE (Analytic Diagonal Estimation):** A non-learning baseline that assumes a diagonal diffusion tensor by setting off-diagonal elements to zero. The diagonal components (D_{xx}, D_{yy}, D_{zz}) are computed from the log-linearized Stejskal-Tanner equation. Specifically, the Apparent Diffusion Coefficient (ADC) derived from each of the three DWIs is assigned to the diagonal element corresponding to the most aligned canonical axis. The final tensor is projected onto the Symmetric Positive Definite (SPD) manifold to ensure physically valid, non-negative eigenvalues.
- **CycleGAN (Zhu et al., 2017):** The architecture consists of two 3D U-Net (Ronneberger et al., 2015) generators and two 3D PatchGAN discriminators, trained with an adversarial loss and an L1 cycle-consistency loss ($\lambda = 10$).
- **Pix2Pix (Isola et al., 2017):** The generator is a 3D U-Net, and the discriminator is a 3D PatchGAN. The training objective is a sum of a vanilla GAN loss and an L1 reconstruction loss ($\lambda_{L1} = 100$).
- **ResViT (Dalmaz et al., 2022):** A hybrid architecture combining a 3D ResNet-style backbone with interleaved Vision Transformer blocks, trained with a composite L1 and adversarial loss.
- **Latent Diffusion (vanilla conditional LDM) (Rombach et al., 2022):** We reuse our frozen Anatomical Autoencoder. The diffusion U-Net is conditioned on the latents of DWI volumes via channel-wise concatenation. Unlike our proposed JET-Diff, this baseline applies a selective denoising strategy, where noise is applied only to the tensor latents, not the full multi-tensor latent.

E USE OF AI TOOLS IN MANUSCRIPT PREPARATION

The authors utilized Google's Gemini Pro to improve the grammar and readability of this manuscript. All content generated by this tool was critically reviewed, fact-checked, and substantially revised by the authors to ensure accuracy and originality. The final responsibility for the content of this paper rests solely with the authors.

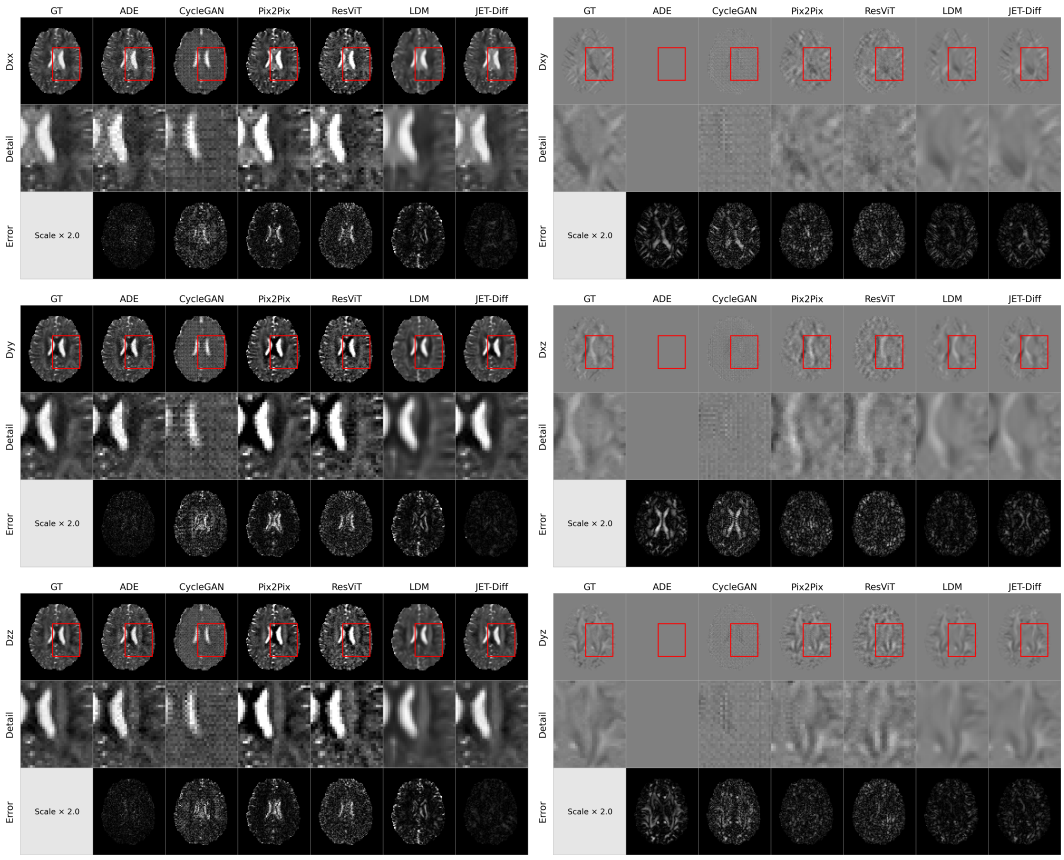


Figure 7: Qualitative comparison of diffusion tensor components. Visualization of the six individual tensor components for the same subject shown in Figure 4. JET-Diff provides a more faithful reconstruction across all components, with significantly reduced noise and artifacts, particularly in the off-diagonal elements (D_{xy}, D_{xz}, D_{yz}).