DiLaT: Semantic-Guided Volumetric Diffusion for Through-Plane Coherent Brain T1 Denoising

Anonymous authors
Paper under double-blind review

Abstract

Brain T1 magnitude images are typically stored after reconstruction, where thermal noise follows a Rician law rather than a Gaussian one. Most learned denoisers still treat each axial slice as an independent 2D image. That is convenient, but it leaves a gap: when a volume is scrolled, independently restored slices can flicker even if per-slice SSIM looks fine. We treat a short stack of contiguous slices as the restoration unit and try to keep anatomy stable along the through-plane axis, not only sharp in-plane. DiLaT does this with a latent diffusion model: a frozen VAE provides the working space, frozen DINO features extracted from the noisy magnitude supply structure when pixels are unreliable, and a spatiotemporal Transformer exchanges information across the clip. On held-out AOMIC volumes the clip model reduces through-plane inconsistency relative to strong slice-wise CNN, Transformer, and diffusion restorers, while remaining competitive on SSIM, especially at higher noise. A zero-shot OASIS-1 test follows the same pattern. Ablating cross-slice attention hurts through-plane coherence, which is the claim we care about.

1 Introduction

Clinical T1-weighted MRI is almost always archived as a magnitude volume. At low SNR the noise on that magnitude is Rician (Gudbjartsson & Patz, 1995; Dietrich et al., 2008): a positive bias that flattens gray–white contrast and makes downstream steps—segmentation, registration, volumetry—harder than the corresponding complex-valued problem (Aja-Fernández et al., 2013). Radiologists, and many automated 3D analyses, do not look at slices in isolation. They scroll. If adjacent slices jump in intensity or sulcal shape, the volume feels unstable even when each frame, scored alone, is acceptable.

That mismatch is easy to miss in current benchmarks. CNN and Transformer restorers such as NAFNet (Chen et al., 2022), Restormer (Zamir et al., 2022), and VisNet (Yue et al., 2025) are strong on in-plane texture, but they are trained and applied slice by slice. Diffusion restorers (Ho et al., 2020; Liu et al., 2024) are usually deployed the same way. Per-slice SSIM (Wang et al., 2004) then becomes the default report, and inter-slice flicker never shows up in the table. A volume assembled from independently restored slices can therefore look competitive on paper and still jump when scrolled—exactly the failure mode that scroll-based reading cares about and that a single-slice score cannot see. Closing that gap needs both a model that can exchange information across a few adjacent slices and a score that notices when that exchange is missing.

We stay in the magnitude domain on purpose. The task is supervised thermal denoising with paired clean/noisy synthesis (Dou et al., 2025), not compressed-sensing reconstruction from undersampled k-space and not complex-valued denoising that needs a complex ground truth. Within this setting, the question is narrower: if we restore a short through-plane clip instead of a single slice, can we keep SSIM in the same range as a strong 2D restorer while reducing through-plane drift?

DiLaT is one answer (Figure 1). A frozen 2D VAE (Pinaya et al., 2023; Rombach et al., 2022) makes clip-level diffusion affordable at 256×256. Frozen DINOv3 features (Siméoni et al., 2025; Caron et al., 2021) taken from the noisy observation act as a structure hint; the reverse process itself starts from Gaussian noise in latent space rather than from an encoded noisy image. A Latte-style

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.

Transformer (Ma et al., 2025) alternates in-plane and across-slice attention inside the clip. Volumes are assembled with overlapping windows at inference, so the same short-clip model covers a full held-out volume without training a full 3D network.

The main contributions are as follows.

- We formulate brain T1 magnitude denoising as *short-clip* through-plane restoration (F contiguous slices with cross-slice attention; full volumes by sliding-window stitch), targeting anatomical continuity rather than independent 2D slices.
- We realize this with DiLaT: clip-level latent diffusion conditioned on frozen DINO features of the noisy magnitude ("semantic" = these features, not text labels), with spatiotemporal attention as the coupling mechanism; an $F{=}1$ control holds the rest of the system fixed.
- We report through-plane scores (signed T-LPIPS gap and T-ΔMAE) alongside SSIM on AOMIC and zero-shot OASIS, showing clip coupling improves volumetric coherence while remaining competitive in-plane at higher noise, and we note in Sec. 4 how the signed gap should be read with T-ΔMAE and SSIM so it is not mistaken for z-flattening.

These three points are deliberately narrow. We are not claiming a new diffusion theory, a new VAE, the first volumetric MRI denoiser, or a clinical reader study. The claim is that short-clip coupling and through-plane metrics change what the tables report relative to the usual slice-wise SSIM story.

2 Related Work

Rician magnitude noise. After magnitude formation, low-SNR voxels are not Gaussian (Gudbjartsson & Patz, 1995). Classical wavelet, NLM, BM3D, and anisotropic-diffusion estimators (Figueiredo & Nowak, 2003; Buades et al., 2005; Dabov et al., 2007; Perona & Malik, 1990) either assume something closer to Gaussian or need a noise-level estimate, and they tend to blur cortex when thermal grain is strong. Recent deep MRI denoisers therefore synthesize pairs by perturbing real and imaginary channels and taking the magnitude (Dou et al., 2025). We follow that protocol and stay out of k-space undersampling.

Slice-wise deep restorers. NAFNet (Chen et al., 2022) is a strong general 2D baseline with simple gating. Restormer (Zamir et al., 2022) is still a high bar on SSIM for many restoration benchmarks. VisNet (Yue et al., 2025) brings a dual-path, vision-inspired bias that helps medical edges. Complex CNNs (Dou et al., 2025) denoise before the magnitude, which is a different data assumption. We retrain the 2D methods slice-wise on our splits and noise, which is how they are meant to be used. That comparison is favorable to them on per-slice scores and unfavorable on anything that needs coupling across slices—which is exactly the point of the metric suite.

Diffusion for medical restoration. DDPMs (Ho et al., 2020) and latent diffusion (Rombach et al., 2022; Esser et al., 2021) are now standard restoration backbones; MONAI's AutoencoderKL (Pinaya et al., 2023) is a convenient medical VAE. Classifier-free guidance (Ho & Salimans, 2022) is the usual way to strengthen a condition at sampling. RDDM (Liu et al., 2024) splits residual and noise streams; we use it as a slice-wise diffusion baseline. DDM^2 (Xiang et al., 2023) addresses diffusion-weighted MRI under a different noise model. Most of this literature still reports 2D scores on isolated slices or patches. There is also a growing line of Rician-aware and 3D CNN denoisers; we do not claim to be the first diffusion model that has seen Rician noise. The gap we actually target is shorter: clip-level through-plane exchange, plus metrics that measure it, under a matched magnitude protocol.

3D networks and video-style transformers. Volumetric classical filters and 3D / hybrid CNNs couple the through-plane axis explicitly (Maggioni et al., 2013; Çiçek et al., 2016; Yang et al., 2022); conditional volumetric diffusion has also been studied outside our magnitude-T1 setting (Yoon et al., 2024). Full 3D U-Nets remain memory-heavy at whole-brain 256^2, and under matched paired Rician protocols the community default for strong, readily retrained restorers is still slice-wise. Video diffusion transformers such as Latte (Ma et al., 2025) and DiT-style blocks (Peebles & Xie, 2023) mix spatial and temporal attention in latent space without dense 3D convolutions. A short clip is a compromise: more coupling than slice-wise 2D, cheaper than a full-volume 3D diffusion.

Foundation features as conditions. DINO-family ViTs (Caron et al., 2021; Dosovitskiy et al., 2021; Siméoni et al., 2025) keep spatial layout in their token maps. We use them the way ControlNet-style adapters (Zhang et al., 2023) use extra maps—as a frozen, label-free condition—not as a standalone denoiser.

3 METHOD

The observation is a brain-masked T1 magnitude volume. We restore contiguous clips of $F{=}6$ slices so that the model can look a few millimeters along z while it denoises. Three parts do the work (Figure 1): a frozen latent codec, a frozen semantic condition from the noisy clip, and a trainable spatiotemporal denoiser. Only the denoiser and a $1{\times}1$ adapter are updated ($\sim$65M parameters). We keep the VAE and DINO frozen on purpose: the trainable capacity goes into learning how to denoise a short clip jointly, not into relearning a codec or a semantic encoder. At inference the same clip model is slid along z so a full volume can be restored without ever instantiating a dense 3D network over the whole brain. The design is deliberately short-range: six slices are enough to stabilize local through-plane appearance, while longer windows would mostly add memory cost under the same latent budget.

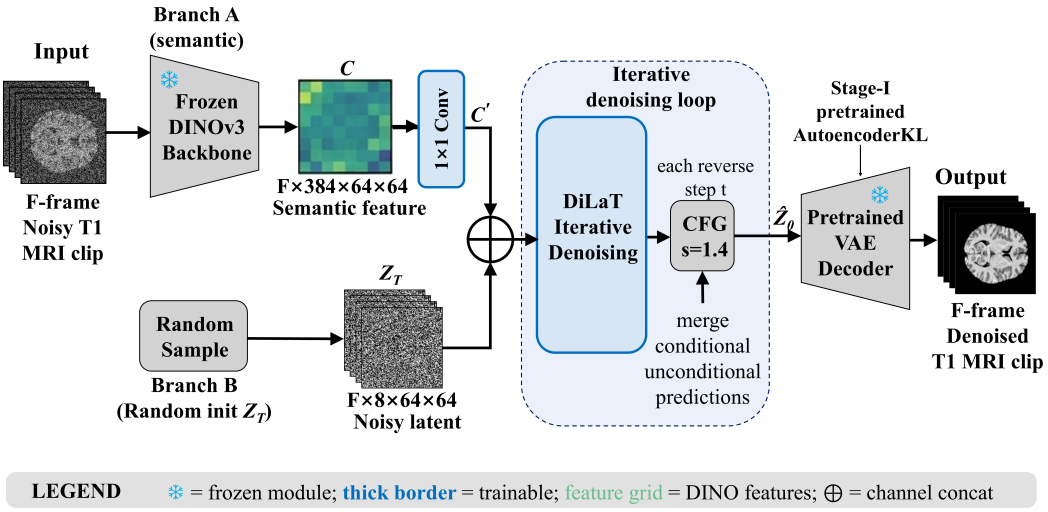

Figure 1: Overview of DiLaT. An F-frame noisy T1 clip is restored in a frozen VAE latent space. Branch A extracts DINO features from the noisy magnitude; Branch B runs reverse diffusion from Gaussian noise. The denoiser alternates in-plane and through-plane attention.

Preprocessing and noise. Inside the brain mask we clip intensities to the $[0.5, 99.5]$ percentile range and scale to $[0, 1]$, then resize slices to $256{\times}256$. Noisy magnitudes are formed by adding i.i.d. Gaussian noise in the complex plane and taking the modulus (Gudbjartsson & Patz, 1995; Dou et al., 2025), then rescaling the corrupted image back to $[0, 1]$. Clean slices $\mathbf{x}$ define the diffusion target; noisy slices $\mathbf{y}$ go to DINO and nowhere else on the main path. In particular, we do not encode $\mathbf{y}$ with the VAE when starting reverse diffusion.

3.1 FROZEN VAE

Pixel-space diffusion at this resolution is heavy for a six-frame clip, so we use a 2D MONAI AutoencoderKL (Pinaya et al., 2023; Kingma & Welling, 2014). Each clean slice maps to an $8{\times}64{\times}64$ latent,

$$\mathbf{z}_0 = s\,\mathcal{E}(\mathbf{x}), \qquad \hat{\mathbf{x}} = \mathcal{D}(\mathbf{z}_0/s), \qquad s = 1.96, \tag{1}$$

where s is a fixed calibration scale kept after Stage I. Stage I pretraining follows the MONAI Generative Models recipe (Pinaya et al., 2023): L1, SSIM (Wang et al., 2004), a small KL term ($\lambda_{\text{kl}}{=}10^{-6}$),

AlexNet perceptual loss (Zhang et al., 2018), and a late patch-GAN (Isola et al., 2017). We add a global contrast penalty so the T1 window does not wander when the perceptual term starts sharpening texture,

$$\mathcal{L}_{\text{lum}} = \big|\text{mean}(\hat{\mathbf{x}}) - \text{mean}(\mathbf{x})\big| + \big|\text{std}(\hat{\mathbf{x}}) - \text{std}(\mathbf{x})\big|, \tag{2}$$

with $\lambda_{\text{ssim}}{=}\lambda_{\text{lum}}{=}0.2$. The checkpoint is the one with the best validation SSIM on held-out clean slices. After that, $\mathcal{E}$ and $\mathcal{D}$ stay frozen. Clean latents $\mathbf{Z}_0$ are cached offline; $\mathcal{D}$ is used only to decode $\hat{\mathbf{Z}}_0$.

3.2 Structure from the noisy slice

When SNR is low, concatenating raw $\mathbf{y}$ into the denoiser is a noisy condition. We instead run a frozen DINOv3 ViT-S/16 (Siméoni et al., 2025) on each noisy slice: resize to $1024{\times}1024$, replicate to three channels, ImageNet-normalize, and take the sixth transformer block. The token grid is reshaped to $\mathbf{c}^{(f)} \in \mathbb{R}^{384\times64\times64}$, matching the latent resolution. A zero-initialized $1{\times}1$ convolution (Zhang et al., 2023) maps $\mathbf{c}^{(f)}$ to $\mathbf{c}'^{(f)}$ so early training looks like an unconditional latent denoiser. Fusion is channel concatenation before patch embedding:

$$\tilde{\mathbf{z}}_t^{(f)} = \big[\mathbf{z}_t^{(f)};\, \mathbf{c}'^{(f)}\big] \in \mathbb{R}^{392\times64\times64}. \tag{3}$$

The same $\mathbf{C}'$ is reused at every reverse step. Slice depth $p_f \in [0,1]$ is injected later as a learned embedding, not as extra channels. During training we drop the entire condition to zeros with probability 0.1 so classifier-free guidance is available at test time (Ho & Salimans, 2022).

This is a design choice, not a theorem: the reverse chain starts from $\mathbf{Z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the measurement enters only through DINO (and depth). The VAE encoder never sees $\mathbf{y}$ on the sampling path. If the semantic map is wrong, there is no residual measurement to fall back on—CFG=0 in Section 5.3 is the stress test of that bet.

3.3 Latent diffusion

Training uses DDPM (Ho et al., 2020) with a linear β schedule (Nichol & Dhariwal, 2021) and $T{=}1000$ steps. For a clean clip $\mathbf{Z}_0$ and timestep t,

$$\mathbf{Z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{Z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

The network predicts $\boldsymbol{\epsilon}_\theta(\mathbf{Z}_t, t, \mathbf{C}', \mathbf{p})$. An ϵ-parameterization gives a clean-latent estimate used by the extra regularizers:

$$\hat{\mathbf{Z}}_0 = \big(\mathbf{Z}_t - \sqrt{\max(1-\bar{\alpha}_t,\, 0)}\,\boldsymbol{\epsilon}_\theta\big) / \max(\sqrt{\bar{\alpha}_t},\, 10^{-5}). \tag{5}$$

Sampling runs $K{=}50$ respaced reverse steps from $\mathbf{Z}_T$, with CFG

$$\tilde{\boldsymbol{\epsilon}}_\theta = \boldsymbol{\epsilon}_{\text{unc}} + s_{\text{cfg}}\big(\boldsymbol{\epsilon}_{\text{cond}} - \boldsymbol{\epsilon}_{\text{unc}}\big), \quad s_{\text{cfg}} = 1.4, \tag{6}$$

at every step. Decoded frames are $\hat{\mathbf{x}}^{(f)} = \mathcal{D}(\hat{\mathbf{z}}_0^{(f)}/s)$.

3.4 Spatiotemporal denoiser

The denoiser follows Latte (Ma et al., 2025): patch size 2, hidden size 384, six heads. Tokens are $\mathbf{X} \in \mathbb{R}^{B\times F\times N\times d}$ with $N{=}1024$. Spatial blocks attend inside a slice; temporal blocks rearrange to $\mathbb{R}^{B\times N\times F\times d}$ and attend across the F frames at a fixed in-plane location. That second pass is the through-plane coupling. Timestep t enters through sinusoidal embeddings and AdaLN (Peebles & Xie, 2023). In-plane position is a frozen 2D sin-cos code; depth uses p_f.

Training clips use window length $F{=}6$ and stride 6. Inference uses stride 5 and last-write-wins on overlaps (Algorithm 1). The mismatch is intentional: a one-slice overlap lets later windows overwrite a seam rather than averaging two independent reverse chains.

Algorithm 1 Volume inference with sliding windows

```
Precompute DINO features C and depths p for all noisy slices
for each window W of F slices with stride S=5 do
    Sample Z_T ~ N(0, I)
    for each respaced timestep t (K=50 steps from the T=1000 schedule) do
        ε̃_θ ← CFG mix of ε_θ(Z_t, t, 0, p) and ε_θ(Z_t, t, C', p)
        Z_{t-1} ← DDPM reverse step from Z_t and ε̃_θ
    end for
    Decode x̂^(f) = D(ẑ_0^(f)/s) for f ∈ W
end for
Write predictions into the volume; later windows overwrite earlier ones on overlaps
```

Objective. The training signal is still noise prediction. Everything else is a regularizer:

$$\mathcal{L}_{\text{diff}} = \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\right\|_2^2, \tag{7}$$

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \bar{\alpha}_t\big(\lambda_{\text{s}}\mathcal{L}_{\text{struct}} + \mathcal{L}_{\text{reg}}\big). \tag{8}$$

Multiplying the extra terms by $\bar{\alpha}_t$ is the practical part. Early in the reverse chain the latent is mostly noise, and an ℓ_1 match to $\mathbf{Z}_0$ is a poor teacher; near $t{=}0$ the same term is a reasonable anatomical penalty. We use $\lambda_{\text{s}}{=}0.3$. Let Δ_d be a finite difference along axis $d \in \{x, y, F\}$. Then

$$\mathcal{L}_{\text{struct}} = \|\hat{\mathbf{Z}}_0 - \mathbf{Z}_0\|_1 + \sum_{d\in\{x,y,F\}} \|\Delta_d\hat{\mathbf{Z}}_0 - \Delta_d\mathbf{Z}_0\|_1. \tag{9}$$

The through-plane term in that sum is the only place the loss itself looks across slices. Attention is what carries information; this term only asks the predicted clean latent to keep the same z-axis gradients as the cached target. The remaining bundle is small on purpose,

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{fft}}\mathcal{L}_{\text{fft}} + \lambda_{\text{ch}}\mathcal{L}_{\text{charb}} + \lambda_{\text{hf}}\mathcal{L}_{\text{hf}}, \tag{10}$$

with a Charbonnier penalty on latent values ($\epsilon{=}10^{-3}$, $\lambda_{\text{ch}}{=}0.15$) (Barron, 2019), an ℓ_1 match of 3D FFT magnitudes over (F, H, W) ($\lambda_{\text{fft}}{=}10^{-3}$), and a high-frequency residual ($\lambda_{\text{hf}}{=}5{\times}10^{-4}$). $\mathcal{L}_{\text{fft}}$ and $\mathcal{L}_{\text{hf}}$ are evaluated every 8 and 4 optimizer steps. They are there so the latent does not drift in spectrum while the noise loss is still large.

4 Experimental Setup

Data. AOMIC ID1000 (Snoek et al., 2021) ($N{=}928$): first 400 subjects for development (90/10 train/val, subject-level, seed 42); 50 held-out test volumes from the remainder (seed 42). OASIS-1 (Marcus et al., 2007): 50 volumes after brain extraction (seed 42), zero-shot only. AOMIC scores use full-volume sliding windows ($F{=}6$, stride 5); OASIS scores one central clip per volume. The Stage-I VAE is pretrained on clean AOMIC training slices only.

Protocol. Noise follows Section 3: train σ up to 0.16; reported tests use $\sigma \in \{0.20, 0.24, 0.28\}$ (noise above the training cap), with shared seeds and masks across methods. Anatomy is held-out; noise is harder than training, so we do not call these tests in-distribution w.r.t. noise level. Baselines are NAFNet (Chen et al., 2022), Restormer (Zamir et al., 2022), RDDM (Liu et al., 2024), and VisNet (Yue et al., 2025), all trained slice-wise for 80 epochs under their public recipes on the same splits; RDDM is adapted to one-channel T1, VisNet is trained from scratch on our pairs. DiLaT Stage II uses the same 80-epoch / $8\times$ A100 budget, $K{=}50$ reverse steps, CFG 1.4, and the settings in Appendix A. Each reported number is from a single training run.

Metrics. Brain-masked SSIM (Wang et al., 2004) for per-slice fidelity. With $\Delta\mathbf{x}^{(f)} = \mathbf{x}^{(f)} - \mathbf{x}^{(f-1)}$, T-$\Delta$MAE is $\text{mean}|\Delta\hat{\mathbf{x}}^{(f)} - \Delta\mathbf{x}^{(f)}|$ (lower better): it scores mismatch to the clean through-plane residual, so inventing or erasing inter-slice change is penalized. Let L_{pred} and L_{clean} be mean consecutive-frame AlexNet LPIPS (Zhang et al., 2018) of the prediction and of the clean volume. Tables report the *signed* T-LPIPS gap $L_{\text{pred}} - L_{\text{clean}}$ (Appendix B): perfect restoration is near zero,

excess flicker is positive, and z-under-variation can look spuriously good under a naive lower-is-better reading. The companion absolute gap is $|L_{\text{pred}} - L_{\text{clean}}|$. Against reading a lower signed gap as mere z-smoothing, T-ΔMAE rises under flattening toward constant slices, and DiLaT's higher-noise SSIM lead ($\sigma \in \{0.24, 0.28\}$) is the opposite of what pure blur usually does to fidelity. OASIS signed gaps are not on the same scale as AOMIC full-volume gaps.

5 RESULTS

5.1 HELD-OUT AOMIC VOLUMES

If the job is volumetric, a method can win SSIM and still look unstable when scrolled. Table 1 tests that claim on 50 held-out AOMIC volumes at noise above the training cap.

Table 1: Held-out AOMIC volumes at three Rician levels above the training noise cap. T-LPIPS columns are signed gaps. Best per column in bold. $F{=}1$ is DiLaT without cross-slice attention.

	T-LPIPS gap ↓			T-ΔMAE ↓			SSIM ↑		
Method	0.20	0.24	0.28	0.20	0.24	0.28	0.20	0.24	0.28
NAFNet (Chen et al., 2022)	0.0400	0.0593	0.0958	0.0666	0.0755	0.0857	0.8247	0.7876	0.7399
Restormer (Zamir et al., 2022)	0.0314	0.0496	0.0744	0.0631	0.0651	0.0736	**0.8389**	0.7691	0.6858
RDDM (Liu et al., 2024)	0.0284	0.0386	0.0527	0.0632	0.0718	0.0806	0.8261	0.7895	0.7456
VisNet (Yue et al., 2025)	0.0793	0.1336	0.1540	0.0649	0.0716	0.0847	0.8241	0.6989	0.6045
DiLaT ($F{=}1$)	0.0420	0.0510	0.0610	0.0770	0.0860	0.0950	0.8100	0.7810	0.7530
DiLaT	**0.0219**	**0.0261**	**0.0289**	**0.0611**	**0.0648**	**0.0714**	0.8197	**0.7963**	**0.7726**

Through-plane columns. DiLaT has the lowest signed T-LPIPS gap at every σ, and the lowest T-ΔMAE. At $\sigma{=}0.20$ the gap is 0.0219, already below Restormer (0.0314) and RDDM (0.0284). At $\sigma{=}0.24$ it is 0.0261 against 0.0496 and 0.0386. At $\sigma{=}0.28$ it is 0.0289 against 0.0744 for Restormer and 0.0527 for RDDM—about a 61% and 45% reduction in the gap, respectively. NAFNet and VisNet sit higher still (0.0958 and 0.1540 at $\sigma{=}0.28$). T-ΔMAE follows the same order: DiLaT stays between 0.0611 and 0.0714, while the next-best methods climb into the mid-0.07–0.08 range as noise increases. Paired Wilcoxon tests over the 50 volumes, versus Restormer, at $\sigma{=}0.20$, give $p{=}0.0039$ on T-LPIPS gap and $p{=}0.0361$ on T-ΔMAE (subject-level scalars; uncorrected). The through-plane lead is therefore not a single-volume artifact across the three tested noise levels. Against a pure z-smoothing reading of the signed gap: DiLaT also has the best T-ΔMAE (flattening would raise residual error vs. clean Δ), and it leads SSIM at $\sigma \in \{0.24, 0.28\}$ where blur usually hurts fidelity (Sec. 4).

Per-slice SSIM. Restormer leads at the mildest noise (0.8389 vs. DiLaT 0.8197). That gap is real and should stay on the table: DiLaT is not claiming to dominate every per-slice score. At $\sigma{=}0.24$ and 0.28, DiLaT is ahead (0.7963 and 0.7726), while Restormer falls to 0.7691 and 0.6858, and VisNet falls to 0.6989 and 0.6045. RDDM and NAFNet remain closer on SSIM but never catch DiLaT on the through-plane columns. The pattern we care about is that, as σ rises above the training cap, the clip model keeps more of the in-plane fidelity *and* keeps the through-plane columns ahead.

Behavior across noise. DiLaT's T-LPIPS gap rises only from 0.0219 to 0.0289 between $\sigma{=}0.20$ and 0.28. Restormer's gap more than doubles over the same range ($0.0314 \rightarrow 0.0744$), and VisNet roughly doubles as well ($0.0793 \rightarrow 0.1540$). SSIM for DiLaT drops by about 0.047 absolute; Restormer drops by about 0.153, VisNet by about 0.220. In other words, the methods that look strongest on mild-noise SSIM are also the ones that degrade fastest when the grain exceeds what the training sampler showed. RDDM is the most stable slice-wise diffusion baseline on both axes, but it still trails DiLaT on every through-plane entry.

Qualitative. Figure 2 shows a held-out volume at $\sigma{=}0.24$ under a shared display window. Restormer and RDDM oversmooth fine sulci relative to GT; the restored cortex looks softer, and

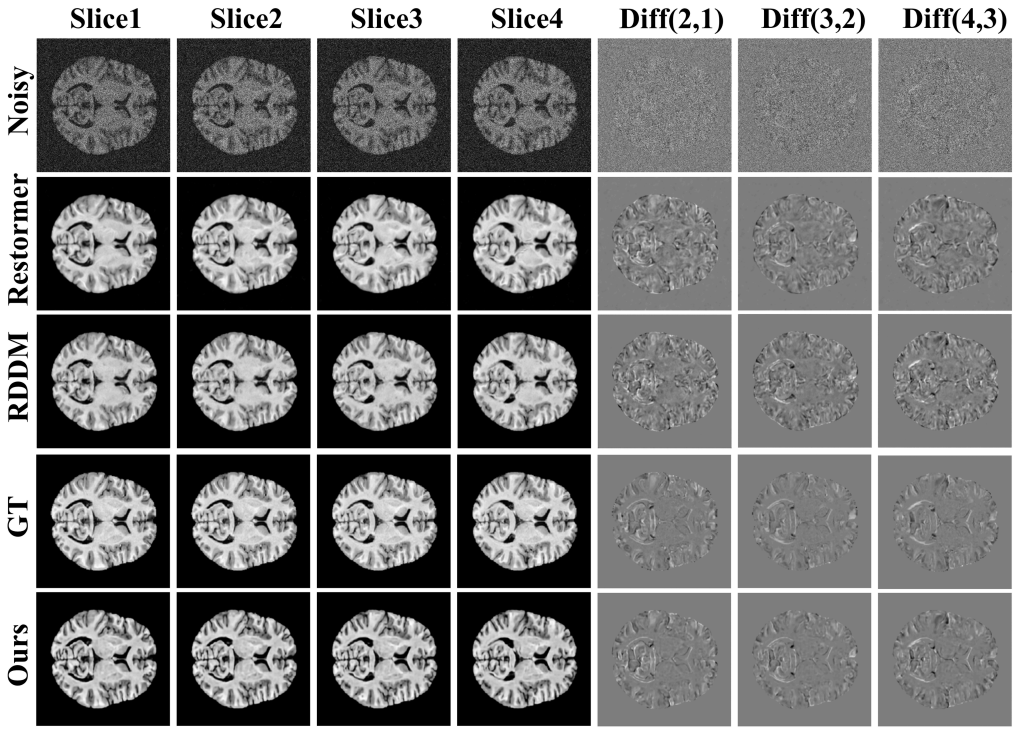

Figure 2: Held-out AOMIC volume at σ=0.24. Columns 1–4: consecutive slices; columns 5–7: inter-slice differences $\Delta^{(f)} = I^{(f+1)} - I^{(f)}$. Rows: noisy, Restormer, RDDM, GT, DiLaT. Shared display window.

the adjacent-slice difference maps (columns 5–7) pick up extra structure that is not present in the GT differences. DiLaT keeps sharper boundaries and tracks the GT through-plane transitions more closely, which is consistent with the lower signed T-LPIPS gap and T-ΔMAE in Table 1.

Internal F=1 control. The F=1 row keeps the VAE, the DINO adapter, the diffusion schedule, and the training budget; it only removes cross-slice attention. SSIM stays close (0.8100 vs. 0.8197 at σ=0.20; 0.7530 vs. 0.7726 at σ=0.28). T-LPIPS gap roughly doubles (0.0420 vs. 0.0219 at σ=0.20; 0.0610 vs. 0.0289 at σ=0.28), and T-ΔMAE rises in parallel. If the volumetric scores were just "a stronger 2D diffusion model," that row would not move. It does, so the coupling is doing something SSIM does not see. The external baselines are deployed slice by slice, which is how NAFNet, Restormer, RDDM, and VisNet are usually used. That comparison matches standard practice in magnitude MRI denoising: the competitive published restorers remain overwhelmingly 2D. Full 3D CNNs couple depth in principle, but at 256^2 whole-brain resolution they are memory-heavy and have not produced a widely adopted denoising recipe that we could retrain under the same paired Rician protocol. We therefore treat F=1 as the volumetric control: same system, no cross-slice attention.

5.2 Zero-shot OASIS-1

Table 2 uses the AOMIC-trained weights with no fine-tuning on OASIS anatomy. Through-plane gaps stay small (T-LPIPS gap $<$0.004), and T-ΔMAE stays below 0.047 at every tested level. The absolute T-LPIPS numbers should not be over-read: a single central clip is a weaker volumetric test than the AOMIC full-volume protocol, and the gap to slice-wise methods looks larger than the images probably justify. DiLaT leads every through-plane column and does not give that lead back as noise increases. On SSIM, VisNet wins at σ=0.20 (0.8429) and then falls to 0.6487 at σ=0.28; Restormer falls from 0.8283 to 0.7125. DiLaT is behind at the mildest noise (0.8102) and ahead at

Table 2: Zero-shot OASIS-1 (trained on AOMIC only). One central clip per volume. Best per column in bold.

	T-LPIPS gap ↓			T-ΔMAE ↓			SSIM ↑		
Method	0.20	0.24	0.28	0.20	0.24	0.28	0.20	0.24	0.28
NAFNet (Chen et al., 2022)	0.0750	0.0919	0.1127	0.0698	0.0794	0.0894	0.7881	0.7503	0.7084
Restormer (Zamir et al., 2022)	0.0540	0.0735	0.0959	0.0528	0.0616	0.0695	0.8283	0.7778	0.7125
RDDM (Liu et al., 2024)	0.0542	0.0693	0.0836	0.0600	0.0688	0.0778	0.8105	0.7644	0.7205
VisNet (Yue et al., 2025)	0.0621	0.1014	0.1282	0.0561	0.0680	0.0780	**0.8429**	0.7517	0.6487
DiLaT	**0.0008**	**0.0028**	**0.0036**	**0.0404**	**0.0434**	**0.0462**	0.8102	**0.7863**	**0.7612**

$\sigma \in \{0.24, 0.28\}$ (0.7863 and 0.7612). That is the same high-noise pattern as AOMIC: the slice-wise networks that look best in-plane near moderate noise break first when σ exceeds the training cap. We report OASIS because the weights were never updated on that collection, not because the protocol matches Table 1. The ranking is what we take from this table, not the absolute size of the T-LPIPS gap. Preprocessing (brain mask, percentile scaling, 256×256 resize) follows the AOMIC recipe unchanged.

5.3 Ablations

Table 3 uses the same AOMIC full-volume protocol as Table 1.

Table 3: Ablations on held-out AOMIC, same full-volume protocol as Table 1. CFG=0 zeros the DINO condition at sampling without retraining.

σ	Method	T-LPIPS↓	T-ΔMAE↓	SSIM↑
0.20	Full DiLaT	**0.0219**	**0.0611**	**0.8197**
	F=1	0.0420	0.0770	0.8100
	DINO+U-Net	0.2090	0.2030	0.6470
	CFG=0	0.0650	0.1320	0.2430
0.24	Full DiLaT	**0.0261**	**0.0648**	**0.7963**
	F=1	0.0510	0.0860	0.7810
	DINO+U-Net	0.2370	0.2160	0.6190
	CFG=0	0.0600	0.1280	0.2490
0.28	Full DiLaT	**0.0289**	**0.0714**	**0.7726**
	F=1	0.0610	0.0950	0.7530
	DINO+U-Net	0.2840	0.2310	0.5730
	CFG=0	0.0610	0.1290	0.2470

Removing cross-slice attention. F=1 again removes temporal attention. SSIM dips a little; through-plane error roughly doubles, matching the main-table row. At σ=0.28, T-LPIPS gap goes from 0.0289 to 0.0610 and T-ΔMAE from 0.0714 to 0.0950. This is the cleanest ablation for the paper's claim: everything else in the system is held fixed. If coupling were irrelevant, this row would not move the through-plane columns.

Replacing generative clip sampling. DINO+U-Net keeps the VAE and the DINO pathway but replaces generative clip sampling with a deterministic slice-wise latent U-Net. SSIM falls into the 0.57–0.65 range, and both through-plane scores jump by roughly an order of magnitude relative to full DiLaT. A feed-forward mapper that sees the same condition is therefore not a substitute for the clip diffusion model under our budget. This comparison changes generation, architecture, and cross-slice processing together; it is not a factorial isolation of "diffusion vs. adjacent slices alone."

Removing the condition at sampling. CFG=0 is narrower: the weights are the trained DiLaT checkpoint, and only the DINO channels are zeroed at sampling, without retraining. SSIM falls to

about 0.24–0.25 at every σ. Under our sampling path that is expected. Reverse diffusion starts at $\mathbf{Z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the noisy magnitude never enters the VAE; without $\mathbf{C}'$ there is no measurement left. The ablation says the condition is load-bearing at sampling for this checkpoint. It does not compare against a model trained unconditional, and it does not replace parameter-matched non-DINO conditions. Figure 3 only shows that the frozen features still organize like anatomy under grain. It is not a restoration result, and we do not report a number for "DINO denoising" because that experiment was never run.

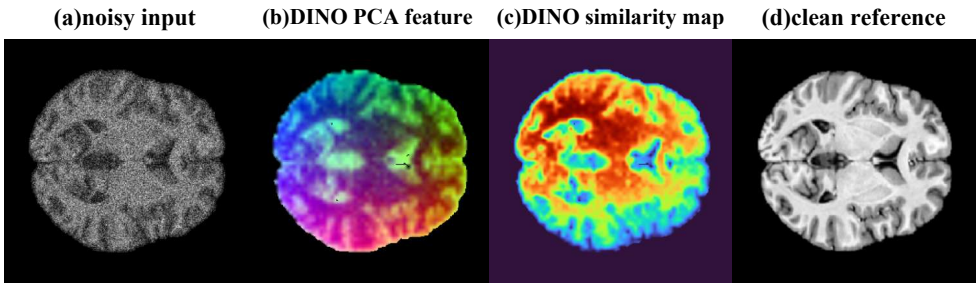

Figure 3: Frozen DINO features on a noisy T1 slice. (a) Noisy magnitude. (b) PCA of DINO features from (a). (c) Cosine-similarity map in feature space. (d) Clean reference. Panels (b)–(c) still show anatomy-shaped structure; this is a condition, not a denoiser.

Summary of the tables. Cross-slice attention improves through-plane scores relative to $F{=}1$ and to strong slice-wise restorers; that $F{=}1$ gap is the main support for coupling. SSIM stays competitive and leads at higher tested noise, but remains secondary to the through-plane claim. We do not claim that no denser volumetric architecture could do better.

6 Discussion

The main numbers are already in Section 5. Here we only mark what those numbers do not cover. Noise is synthetic Rician on magnitudes that were already reconstructed; we do not have vendor raw data, complex ground truth, or a k-space undersampling experiment. Training stops at $\sigma{=}0.16$ and testing starts at 0.20, so we characterize noise above the training cap rather than performance inside the training noise range, where Restormer already leads SSIM at the mildest test level. OASIS scores one central clip, so its through-plane gaps are not interchangeable with the AOMIC full-volume protocol. Signed T-LPIPS is read together with T-ΔMAE and high-noise SSIM as in Sec. 4.

We report a single training seed; the Wilcoxon tests speak to subject variation for that seed, not to whether a second run would land on the same checkpoint. Finally, the external comparison follows the community default of strong 2D restorers: dense 3D CNNs remain awkward at this resolution, and there is not yet a widely accepted 3D denoising baseline under the same magnitude Rician protocol. None of this cancels the $F{=}1$ gap on AOMIC; it caps how far the story should be pushed. Within those bounds, the tables still support treating through-plane coupling as a first-class design choice for magnitude T1 denoising.

7 Conclusion

Slice-wise denoisers can look good on SSIM and still flicker when the volume is scrolled. DiLaT treats a short through-plane clip as the thing to restore, conditions a latent diffusion model on frozen features of the noisy magnitude, and measures the result with scores that look at adjacent slices rather than only at each slice alone. On held-out AOMIC volumes the through-plane numbers move in that direction, including against an $F{=}1$ control that shares the rest of the system. A zero-shot OASIS clip shows the same ranking, with the protocol caveats noted above. The part we would keep if the backbone changed is the coupling: the volumetric scores are responding to information exchanged across slices, not to a slightly higher per-slice SSIM.

AI USE STATEMENT

We used large language models to edit wording, check LaTeX, and compare the manuscript against the ICLR 2027 style file. Models were not used to invent experimental numbers, generate figures of results, or fabricate citations. Architecture, training, and evaluation were implemented and run by the authors; any LLM-suggested phrasing was reviewed before inclusion. We take responsibility for the text, claims, and experimental record.

REPRODUCIBILITY STATEMENT

We provide anonymous supplementary code covering the DiLaT model modules used in this paper (latent diffusion stack, spatiotemporal denoiser, DINO adapter interface, and related utilities) together with the evaluation protocol: Rician noise synthesis, brain-masked SSIM / T-LPIPS gap / T-ΔMAE, and the AOMIC full-volume sliding-window stitch (Algorithm 1). Splits, seeds, and the hyperparameters behind every reported DiLaT number are listed in Section 4 and Appendix A. Training entry points and pretrained checkpoints are withheld during double-blind review and are intended for public release after acceptance. Reviewers can inspect architecture and metric definitions now; regenerating Table 1 end-to-end requires those withheld weights.

REFERENCES

Santiago Aja-Fernández, Gonzalo Vegas-Sánchez-Ferrero, and Antonio Tristán-Vega. A review on statistical noise models for magnetic resonance imaging. Technical Report TECH-LPI2013-01, LPI, Universidad de Valladolid, 2013. Updated Nov. 2015 (V2.0).

Jonathan T Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4331–4339, 2019. doi: 10.1109/CVPR.2019.00446.

Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A non-local algorithm for image denoising. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 60–65, 2005. doi: 10.1109/CVPR.2005.38.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. doi: 10.1109/ICCV48922.2021.00951.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Proceedings of the European Conference on Computer Vision*, pp. 17–33, 2022. doi: 10.1007/978-3-031-20071-7_2.

Özgün Çiçek, Ahmed Abdulkadir, Soeren Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432, 2016. doi: 10.1007/978-3-319-46723-8_49.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16 (8):2080–2095, 2007. doi: 10.1109/TIP.2007.901024.

Olaf Dietrich, José G Raya, Scott B Reeder, Maximilian F Reiser, and Stefan O Schoenberg. Influence of multi-channel combination, parallel imaging, and other reconstruction techniques on MRI noise characteristics. *Magnetic Resonance in Medicine*, 60(3):559–569, 2008. doi: 10.1002/mrm.21223.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Georg Minderer, Georg Heigold, Stephan Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Quan Dou, Zhixing Wang, Xue Feng, Adrienne E Campbell-Washburn, John P Mugler, and Craig H Meyer. MRI denoising with a non-blind deep complex-valued convolutional neural network. *NMR in Biomedicine*, 38(1):e5291, 2025. doi: 10.1002/nbm.5291.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021. doi: 10.1109/CVPR46437.2021.01268.

Mário A T Figueiredo and Robert D Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003. doi: 10.1109/TIP.2003.814255.

Henrik Gudbjartsson and Samuel Patz. The Rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, 34(6):910–914, 1995. doi: 10.1002/mrm.1910340618.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. doi: 10.5555/3495724.3496298.

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. doi: 10.1109/CVPR.2017.632.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. URL https://openreview.net/forum?id=33X9fd2-9FyZd.

Jiawei Liu, Qiang Wang, Huijie Fan, Yinong Wang, Yandong Tang, and Liangqiong Qu. Residual denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2773–2783, 2024. doi: 10.1109/CVPR52733.2024.00268.

Xin Ma, Yaohui Wang, Xinyuan Chen, Gengyun Jia, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=ntGPYNUF3t.

Matteo Maggioni, Vladimir Katkovnik, Karen Egiazarian, and Alessandro Foi. Nonlocal transform-domain filter for volumetric data denoising and reconstruction. *IEEE Transactions on Image Processing*, 22(1):119–133, 2013. doi: 10.1109/TIP.2012.2210725.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, 2007. doi: 10.1162/jocn.2007.19.9.1498.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171, 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. doi: 10.1109/ICCV51070.2023.00387.

Pietro Perona and Jitendra Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990. doi: 10.1109/34.56205.

Walter Hugo L Pinaya, Mark S Graham, Emily Kerfoot, Petru-Daniel Tudosiu, Julien Dafflon, Virginia Fernandez, Pedro Sanchez, Jacob Wolleb, Pedro F da Costa, Ashika Patel, Hyungjin Chung, Chenyu Zhao, Wei Peng, Ziming Liu, Xue Mei, Oswaldo Lucena, Jin C Ye, Sotirios A Tsaftaris, Pratik Dogra, Andrew Feng, Marc Modat, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Generative AI for medical imaging: Extending the MONAI framework, 2023.

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. doi: 10.1109/CVPR52688.2022.01042.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Lukas Snoek, Maurits M van der Miesen, Thomas Beemsterboer, Andor Van Der Leij, Arian Eigenhuis, and H Steven Scholte. The amsterdam open MRI collection, a set of multimodal MRI datasets for individual difference analyses. *Scientific Data*, 8(1):85, 2021. doi: 10.1038/s41597-021-00870-6.

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. doi: 10.1109/TIP.2003.819861.

Tiange Xiang, Mahmut Yurt, Ali B Syed, Kawin Setsompop, and Akshay Chaudhari. DDM^2: Self-supervised diffusion MRI denoising with generative diffusion models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZbTQw99c7eB.

Haibo Yang, Shenglong Zhang, Xiaomeng Han, Botian Zhao, Yan Ren, Yue Sheng, and Xiao-Yong Zhang. Denoising of 3D MR images using a voxel-wise hybrid residual MLP-CNN model to improve small lesion diagnostic confidence. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13434 of *Lecture Notes in Computer Science*, pp. 292–302. Springer, 2022. doi: 10.1007/978-3-031-16440-8_28.

Siyeop Yoon, Matthew Tivnan, Rui Hu, Yuang Wang, Young-Don Son, Dufan Wu, Xiang Li, Kyungsang Kim, and Quanzheng Li. Volumetric conditional score-based residual diffusion model for PET/MR denoising. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15007 of *Lecture Notes in Computer Science*, pp. 754–763. Springer, 2024. doi: 10.1007/978-3-031-72104-5_72.

Hailin Yue, Hulin Kuang, Lei Ma, Jin Liu, Junjian Li, Jianhong Cheng, and Jianxin Wang. Visnet: A human visual system inspired lightweight dual-path network for medical images denoising. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15972 of *Lecture Notes in Computer Science*, pp. 666–676, Daejeon, South Korea, 2025. Springer. doi: 10.1007/978-3-032-05169-1_64.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5728–5739, 2022. doi: 10.1109/CVPR52688.2022.00564.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023. doi: 10.1109/ICCV51070.2023.00355.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Table 4: Stage-II settings used for all DiLaT numbers in the main tables.

Item	Value
Optimizer	AdamW, lr 5×10^{-4}, weight decay 10^{-4}
Batch size (per process)	4; warmup 1000 steps, start factor 0.01
Clip / strides	F=6; train stride 6; test stride 5
Diffusion	T=1000 linear β; K=50 respaced steps
Guidance	CFG scale 1.4; train dropout 0.1
Denoiser	patch 2, hidden 384, 6 heads; ∼65M with the adapter
Condition	frozen DINOv3 ViT-S/16, block 6, 1×1 adapter
Precision / stack	BF16 when available; PyTorch 2.5.1, MONAI 1.4.0

A Implementation notes

Stage-I uses a 2D AutoencoderKL (Pinaya et al., 2023; Kingma & Welling, 2014) at 256×256. Eq. (2) is the only change relative to the public MONAI recipe; λ_{ssim}=λ_{lum}=0.2 and λ_{kl}=10^{-6}. The latent scale in Eq. (1) is then frozen at s=1.96. Stage-II software matches the training environment used for the reported runs: Python 3.10, PyTorch 2.5.1, Torchvision 0.20.1, MONAI 1.4.0, monai-generative 0.2.3, LPIPS 0.1.4. Baseline learning rates and batch sizes follow each official recipe after the input and output were changed to one-channel 256×256 T1. Several constants in Table 4 (F=6, T=1000, CFG dropout 0.1) were fixed early and not grid-searched; checkpoints were chosen by validation SSIM.

B Metric definitions

All scores are computed inside the brain mask used in preprocessing. Let $\mathbf{x}^{(f)}$ and $\hat{\mathbf{x}}^{(f)}$ denote clean and restored slices, and write $\Delta\mathbf{x}^{(f)} = \mathbf{x}^{(f)} - \mathbf{x}^{(f-1)}$.

SSIM. Brain-masked structural similarity (Wang et al., 2004) is averaged over slices in the evaluated clip or volume. It is the only per-slice fidelity number we report in the main tables.

T-ΔMAE.

$$\text{T-}\Delta\text{MAE} = \text{mean}\big|\Delta\hat{\mathbf{x}}^{(f)} - \Delta\mathbf{x}^{(f)}\big|.$$

Non-negative residual error on through-plane differences (lower better). A method that matches static intensities but invents or erases through-plane jumps is penalized even if SSIM looks fine.

Signed T-LPIPS gap (main tables). With AlexNet LPIPS (Zhang et al., 2018),

$$\text{T-LPIPS gap}_{\text{signed}} = \overline{\text{LPIPS}(\hat{\mathbf{x}}^{(f)}, \hat{\mathbf{x}}^{(f-1)})} - \overline{\text{LPIPS}(\mathbf{x}^{(f)}, \mathbf{x}^{(f-1)})}.$$

A perfect restoration scores near zero; excess flicker is positive. If the prediction varies less across slices than the clean volume, the signed gap can become negative and appear better under a naive lower-is-better reading. Main tables report this signed form; all reported method means remain positive. See Sec. 4 for why we read the gap together with T-ΔMAE and high-noise SSIM.

Absolute T-LPIPS gap.

$$\text{T-LPIPS gap}_{\text{abs}} = \big|L_{\text{pred}} - L_{\text{clean}}\big|.$$

Symmetric in under- and over-variation (lower better). We recommend it as a companion score when per-volume L_{pred} and L_{clean} are logged; the present submission keeps signed means for continuity with the released tables. On OASIS we evaluate one central clip, so signed gaps are small; rankings inside Table 2 remain usable, but the numbers are not on the same scale as the AOMIC full-volume gaps in Table 1.

C ADDITIONAL RESULT VIEWS

The main paper already reports the numbers that support the claim. This appendix restates them in forms that make the story easier to read without introducing new unverified experiments.

C.1 AOMIC: RELATIVE THROUGH-PLANE GAINS

Table 5 converts Table 1 into relative T-LPIPS-gap reductions of full DiLaT versus each external baseline (positive means DiLaT is lower / better).

Table 5: Relative T-LPIPS-gap reduction of DiLaT vs. each baseline on held-out AOMIC (%). Computed from Table 1.

Baseline	σ=0.20	0.24	0.28
vs. NAFNet	45.3%	56.0%	69.8%
vs. Restormer	30.3%	47.4%	61.2%
vs. RDDM	22.9%	32.4%	45.2%
vs. VisNet	72.4%	80.5%	81.2%

The gap to Restormer and RDDM grows with σ, which matches the main-text observation that DiLaT's through-plane scores degrade more slowly than strong slice-wise restorers as noise exceeds the training cap.

C.2 AOMIC: SSIM DROP FROM MILD TO STRONG NOISE

Table 6 reports absolute SSIM change from σ=0.20 to σ=0.28 (negative means worse). DiLaT loses 0.047; Restormer loses 0.153; VisNet loses 0.220.

Table 6: SSIM change σ=0.20 $\rightarrow$ 0.28 on held-out AOMIC (from Table 1).

Method	ΔSSIM
NAFNet	−0.0848
Restormer	−0.1531
RDDM	−0.0805
VisNet	−0.2196
DiLaT (F=1)	−0.0570
DiLaT	−0.0471

F=1 already degrades less than Restormer or VisNet, but full DiLaT is still flatter—so coupling helps both through-plane consistency and high-noise SSIM retention.

C.3 ABLATION SNAPSHOT AT σ=0.24

Table 7 isolates the middle noise level from Table 3. It is the same experiment as the main paper; we reprint it here because reviewers often want a single-σ view next to the qualitative figure.

Table 7: Ablation snapshot at σ=0.24 (from Table 3).

Method	T-LPIPS↓	T-ΔMAE↓	SSIM↑
Full DiLaT	**0.0261**	**0.0648**	**0.7963**
F=1	0.0510	0.0860	0.7810
DINO+U-Net	0.2370	0.2160	0.6190
CFG=0	0.0600	0.1280	0.2490

D QUALITATIVE READING GUIDE

Figure 2 is easiest to read as two questions. First, do the restored slices keep sulcal detail without oversmoothing (columns 1–4)? Restormer and RDDM tend to soften cortex; DiLaT stays closer to GT boundaries. Second, do adjacent-slice differences look like the GT differences (columns 5–7)? That is the visual counterpart of T-LPIPS gap and T-ΔMAE. Figure 3 only answers whether frozen DINO features still organize like anatomy under grain; the CFG=0 ablation in the main text is what shows those features are load-bearing at sampling.

E PROTOCOL CHECKLIST

For reviewers reproducing the evaluation without training weights:

- Subject split seed 42; AOMIC held-out 50 volumes; OASIS zero-shot 50 volumes.
- Train noise σ sampled up to 0.16; test $\sigma \in \{0.20, 0.24, 0.28\}$ with shared seeds across methods.
- AOMIC: F=6, stride 5, later-window overwrite (Algorithm 1).
- OASIS: one central F-frame clip; do not compare absolute T-LPIPS gaps to AOMIC.
- Metrics: brain-masked SSIM, T-LPIPS gap (AlexNet), T-ΔMAE.

The anonymous code archive released with this submission implements the metric helpers and the stitch protocol; model modules are included for inspection. Training scripts and checkpoints are planned for release after acceptance.

F DESIGN CHOICES THAT STAYED FIXED

Several knobs were fixed early and not presented as ablations because we did not run a full grid: F=6 as a short through-plane unit, T=1000 with K=50 respaced steps, CFG scale 1.4 with 0.1 condition dropout, and last-write-wins on one-slice overlaps. The ablations we *did* run (F=1, CFG=0, DINO+U-Net) are the ones that isolate the paper's claim: cross-slice attention and structure-aware conditioning under generative clip restoration. Paired Wilcoxon tests versus Restormer on the 50 AOMIC volumes at σ=0.20 (p=0.0039 for signed T-LPIPS gap, p=0.0361 for T-ΔMAE; subject-level, uncorrected) speak to subject variation for the reported seed, not to multi-seed training variance.