Wavelet whitened patch-based diffusion prior for velocity models

Ege Cirakman[*1], *Huseyin Tuna Erdinc*[*2], *Felix J. Herrmann*[2]
[1]*Istanbul Technical University,* [2]*Georgia Institute of Technology*

SUMMARY

Many physical properties, such as subsurface velocity models, exhibit fractal-like $1/f^{\alpha}$ power spectra with $\alpha > 1$, implying a power-law decay of spectral energy with frequency, where low-frequency components dominate and high-frequency contributions decay polynomially. This induces long-range spatial correlations and complicates the estimation of statistical quantities, most notably yielding ill-conditioned score functions central to score-based generative models. Recent patch-based training strategies have been introduced to mitigate the challenges of high-dimensional learning; however, for data with strong long-range correlations, patching further exacerbates ill-posedness in score estimation. To address this challenge, we propose a scale- and orientation-normalized invertible wavelet transform tailored for patch-based training of score-based generative models. Additionally, we introduce a sampling correction that compensates for numerical inconsistencies induced by the multiscale transform during inference. Empirical results show that, relative to baseline patch-based models, the proposed approach achieves an average improvement of approximately 20% in power spectral fidelity, while reducing training and inference time by 20% and 50%, respectively.

INTRODUCTION

Subsurface velocity models can be characterized by their $1/f^{\alpha}$ decay in power spectra (Wainwright and Simoncelli, 2000; Turcotte, 1997). This behavior reflects the geology of sedimentary basins, where broad low-frequency trends encode regional stratification, while high-frequency components capture interfaces and lateral heterogeneity. Such structure arises from the hierarchical and multiscale nature of geological processes—sedimentation, compaction, diagenesis, and tectonic deformation—which operate across a wide range of spatial scales. These mechanisms induce approximate fractal scaling (Frankel and Clayton, 1986; Herrmann, 1997), yielding long-range spatial correlations and self-similar organization in subsurface properties.

Recent advances in generative modeling have enabled the estimation of complex distributions over Earth properties, such as velocity fields. Among these approaches, score-based generative models learn the score function during training and generate samples at inference via iterative refinement. A key factor governing both training stability and sampling convergence is the conditioning of the score covariance. Distributions exhibiting strong long-range correlations are known to induce ill-conditioned covariance structures, thereby degrading both learning efficiency and sampling robustness.(Mallat, 2009).

In Bayesian seismic inversion settings—where the objective is to infer velocity from observations such as Reverse Time Migration (RTM) images (Erdinc et al., 2025) or common image gathers (Yin et al., 2024)—this issue becomes particularly acute due to the high dimensionality of the parameter space. The combination of large-scale inference and long-range correlations renders conventional approaches computationally burdensome and statistically ill-posed. Prior work (Cirakman et al., 2025; Guth et al., 2022) addresses this challenge by decomposing the problem across wavelet scales and enforcing approximate statistical whitening via scale-wise normalization. These multiscale, cascaded architectures—progressing from coarse to fine resolution—not only improve conditioning but also yield substantial computational gains; for example, (Cirakman et al., 2025) reports reductions of approximately 50% in GPU memory usage and 73% in posterior sampling time.

More recently, the field has shifted toward increasingly high-dimensional settings, such as 3D volumes and panoramic 2D domains, motivating the use of patch-based methods for scalability. Patch diffusion (Hu et al., 2024) enables full-image score estimation from localized patch computations through positional encodings, significantly reducing memory requirements during training.However, patch-wise modeling disrupts long-range dependencies, further exacerbating the ill-conditioning of the score covariance (Orozco et al., 2024; Zeng et al., 2025).

Motivated by these challenges, we propose an integration of normalized multiscale wavelet whitening within patch-based score modeling frameworks to restore statistical conditioning while retaining computational scalability.

THEORY AND METHOD

Score-based and patch-based diffusion models

A score-based diffusion model learns the gradient of the log-density, $\nabla_{\mathbf{x}} \log p(\mathbf{x})$, known as the score function, via a denoising objective (Song et al., 2021). A neural network $D_\theta(\mathbf{x}+\varepsilon, \sigma_t)$ is trained to recover a clean image $\mathbf{x}$ from its noisy version at noise level σ_t; the score follows from Tweedie's formula as $s_\theta = (D_\theta - \mathbf{x})/\sigma_t^2$. Sampling iterates the learned score from pure noise through a reverse-time stochastic differential equation (Anderson, 1982). When the data covariance is ill-conditioned—as in velocity models with $1/f$ spectra—the score magnitude varies by orders of magnitude across scales, slowing both training and sampling (Guth et al., 2022).

PaDIS (Hu et al., 2024) sidesteps the memory bottleneck by expressing the full-image score as a sum of patch-level scores. Each patch carries a positional encoding so the network learns global structure from local crops. The image is zero-padded and partitioned into non-overlapping $P \times P$ patches under random shifts; averaging over all shifts yields a consistent estimator of $\nabla \log p(\mathbf{x})$, making PaDIS the only current route toward three-dimensional diffusion priors.

Wavelet preconditioning and boundary-aware sampling for PaDIS

Diagonal Wavelet Whitening

Let $\mathbf{x} \in \mathbb{R}^{n_x \times n_z}$ denote a velocity model with n_x horizontal and n_z depth grid points. A J-level discrete wavelet transform $\mathrm{WT}_J^{\mathscr{B},\gamma}$ with the Daubechies-2 wavelet (Daubechies, 1992) decomposes the image into one low-pass subband per level and three oriented detail subbands. We choose db2 over Haar because its two vanishing moments annihilate linear trends, producing sparser coefficients for piecewise-smooth geology (Herrmann, 1997). We use periodization as the boundary convention ($\mathscr{B}$), so the transform introduces no boundary artifacts. At each location u, the three detail subbands are stacked into $\overline{\mathbf{x}}_j(u) \in \mathbb{R}^3$: LH (low-pass rows, high-pass columns—horizontal edges, dominant in layered models), HL (high-pass rows, low-pass columns—vertical edges), and HH (high-pass both—diagonal structure). Recursive decomposition of the low-pass yields coarse coefficients $\mathbf{x}_J$ and details $\{\overline{\mathbf{x}}_j\}_{j=1}^J$ at scales $j = 1, \ldots, J$ (finest to coarsest).

Following WSGM (Guth et al., 2022), we normalize each detail subband by $\gamma_j = E_j^{1/2}$, the root-mean-square detail energy at scale j over the training set, so that $\overline{\mathbf{x}}_j = \gamma_j^{-1}\,\overline{\mathbf{x}}_j^{\mathrm{raw}}$ has unit expected energy. The covariance $\Sigma_j^{HH} \in \mathbb{R}^{3\times 3}$ of the normalized detail vector has diagonal entries giving per-orientation variance. In velocity models, the LH variance dominates HL and HH by orders of magnitude, reflecting the strong horizontal anisotropy of velocity models.

The diagonal standardization matrix $\mathbf{D}_j = (\mathrm{diag}(\Sigma_j^{HH}) + \varepsilon^2 \mathbf{I}_3)^{-1/2}$, where ε^2 prevents division by near-zero variance, yields whitened details

$$\widetilde{\overline{\mathbf{x}}}_j(u) = \mathbf{D}_j\big(\overline{\mathbf{x}}_j(u) - \mu_j^H\big), \tag{1}$$

with $\mu_j^H \in \mathbb{R}^3$ the dataset mean detail vector. The coarsest low-pass is standardized as $\widetilde{\mathbf{x}}_J = (\mathbf{x}_J - \mu_J^L)/\sqrt{\Sigma_J^{LL} + \varepsilon^2}$, with μ_J^L and Σ_J^{LL} its scalar mean and variance. We denote the full per-scale whitening operator by B and reassemble via $\mathbf{w} = \mathrm{IWT}_J^{\mathscr{B},\gamma}(\widetilde{\mathbf{x}}_J, \{\widetilde{\overline{\mathbf{x}}}_j\}_{j=1}^J)$.

A global affine map $G(\mathbf{w}) = (\mathbf{w} - \mu_{\mathrm{img}})/(\sigma_{\mathrm{img}} + \delta)$—with μ_{img} and σ_{img} the dataset mean and standard deviation of the whitened images and δ a small stabilization constant—sets zero mean and unit variance, matching the EDM (Karras et al., 2022) noise schedule. The calibrated $\mathbf{u} = G(\mathbf{w})$ has heavy tails due to wavelet coefficient kurtosis; soft compression $\mathbf{z} = C\tanh(\mathbf{u}/C)$ with C=4 bounds the data to $[-4, +4]$. The inverse $\mathbf{u} = C\,\mathrm{arctanh}(\mathbf{z}/C)$ is applied at dewhitening. The full chain is exactly invertible at machine precision.

The wavelet transform and diagonal standardization are both separable, so the whitening applies identically to 2D slices and to 3D volumes with linear cost scaling and no architectural changes to the patch-based model.

Boundary-aware sampling

Standard PaDIS assembles full-image predictions by hard write-back within each non-overlapping partition, then averages over a fixed set of shifts. With few diffusion steps, the same boundaries recur at every iteration, producing visible discontinuities that break geological continuity. We replace this procedure with three modifications. First, we draw partition offsets at each step from a Halton low-discrepancy sequence (Halton, 1960; Niederreiter, 1992)—a deterministic scheme that fills the shift space more uniformly than pseudorandom draws, so consecutive steps never share boundary locations. Second, we replace hard write-back with weighted fusion: each patch prediction is multiplied by a separable cosine window $W_{ij} = \omega(i)\,\omega(j)$, where ω is a raised-cosine taper (Harris, 1978) that downweights pixels near patch edges while preserving full weight at centers. The fused prediction is the pointwise weighted average over all shifts τ and patches r. Third, for border patches the taper is flattened on the border-facing side, preventing near-zero weights from suppressing edge predictions.

The sampler extends to 3D volumes without modification. In three dimensions, a third Halton base is added so that offsets along each axis follow independent, non-repeating patterns, and the cosine window factors as $W_{ijk} = \omega(i)\,\omega(j)\,\omega(k)$—one extra multiplication per voxel, no additional design.

Training objective

Given a velocity model $\mathbf{x}$, the whitening pipeline produces $\mathbf{z} = C\tanh(G(B(\mathbf{x}))/C)$, where B applies the per-scale wavelet whitening and inverse transform, G the global calibration, and C=4 the soft compression. The preconditioned image $\mathbf{z}$ is zero-padded and partitioned into $P{\times}P$ patches; each patch $\mathbf{z}_r$ is paired with a two-channel positional encoding $\mathbf{p}_r$ that carries its absolute spatial coordinates on the canvas.

A single denoiser $D_\theta(\mathbf{z}_r + \varepsilon, \mathbf{p}_r, \sigma_t)$ is trained to recover $\mathbf{z}_r$ from its noisy version at noise level σ_t; the patch-level score follows from Tweedie's formula as $s_\theta = (D_\theta - (\mathbf{z}_r + \varepsilon))/\sigma_t^2$. Following EDM (Karras et al., 2022), the loss is

$$\mathscr{L}(\theta) = \mathbb{E}_{\sigma_t}\,\mathbb{E}_{\mathbf{z}}\,\mathbb{E}_{\varepsilon}\big[\lambda(\sigma_t)\,\|D_\theta(\mathbf{z}_r + \varepsilon, \mathbf{p}_r, \sigma_t) - \mathbf{z}_r\|_2^2\big], \tag{2}$$

where $\varepsilon \sim \mathscr{N}(\mathbf{0}, \sigma_t^2\mathbf{I})$ and $\lambda(\sigma_t) = 1/\sigma_t^2$. Unlike WSGM (Guth et al., 2022), which trains a separate network per wavelet scale with cascaded conditioning on parent coefficients, our whitening equalizes subband statistics so that a single network handles all scales. The positional encoding $\mathbf{p}_r$ replaces explicit coarse-scale conditioning by providing depth-dependent context to each patch. Figure 1 summarizes the full pipeline.

EXPERIMENTS AND RESULTS

Dataset and training configuration

We use a synthetic 3D Earth model derived from the Compass model, representative of North Sea geological formations (Jones et al., 2012). The training set consists of 3 000 two-dimensional velocity slices (256×256 pixels, 6.25 m spacing, covering 3.2 × 3.2 km), split into 2 500 training and 500 validation samples.

Each velocity image is whitened with Diagonal Wavelet Whitening (db2, J=6, C=4). We train a DDPM (Karras et al., 2022) on 64×64 patches (P=64) with batch size 4, learning rate 10^{-4}, and the multi-scale crop strategy of PaDIS (Hu et al., 2024). Training runs for 900 epochs (whitened) on a single GPU.

Wavelet preconditioning and boundary-aware sampling for PaDIS

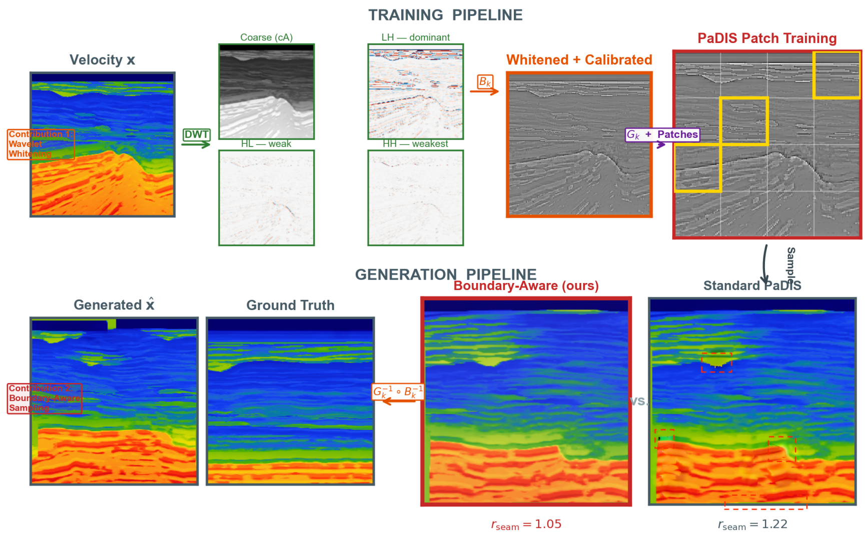

Figure 1: Proposed pipeline. Top: wavelet transform, diagonal whitening (B), calibration (G), and patch extraction for PaDIS training. Bottom: boundary-aware sampling replaces hard assembly; exact dewhitening ($G^{-1} \circ B^{-1}$) recovers the velocity model.

Spectral conditioning

The $1/f^{\alpha}$ spectrum of velocity models concentrates energy at coarse scales, producing a radial spectral slope of -2.8 that spans five decades in power (Figure 2a, c). Diagonal Wavelet Whitening reduces this slope to -0.3 (Figure 2e), equalizing the score-matching target across spatial frequencies. The 2D spectrum confirms this: the strong vertical bar in the original (Figure 2c), reflecting horizontal-layer energy, spreads into a more uniform distribution after whitening (Figure 2d). The transform preserves machine-precision invertibility, so no information is lost.

The practical consequence is faster training. The whitened model converges at 60 epochs/hour, compared to 40 epochs/hour for the unwhitened model—a 50% wall-clock speedup under identical architecture and hyperparameters.

Directional spectral fidelity

We evaluate fidelity per wavelet orientation using directional power spectra. Given a velocity model $\mathbf{x}$ with 2D power spectrum $S(f_1, f_2) = |\mathscr{F}\{\mathbf{x}\}(f_1, f_2)|^2$, the directional spectrum $P_d(f) = \int S(f_d, f_\perp)\, df_\perp$ integrates out the frequency axis perpendicular to direction d, yielding a one-dimensional power curve. LH measures horizontal-layer energy (vertical frequency), HL lateral variations (horizontal frequency), and HH diagonal detail. Figure 3 compares these curves for the ground truth, PaDIS, and the proposed whitening. In LH, whitening reduces the slope mismatch from 0.11 to 0.01 (-1.96 vs. GT -1.95). The subdominant orientations show larger corrections: in HL, the mismatch drops from 0.41 to 0.25 (-2.00 vs. GT -2.25); in HH, from 0.43 to 0.05 (-3.06 vs. GT -3.01). Whitening improves fidelity most in the orientations where PaDIS under-invests network capacity.

Figure 4 compares generated velocity models at equal training compute. PaDIS (a) loses fine horizontal layering and shows faint seam artifacts at patch boundaries. With Diagonal Wavelet Whitening and boundary-aware sampling (b), the model recovers lateral velocity variations and produces smoother transitions across boundaries.

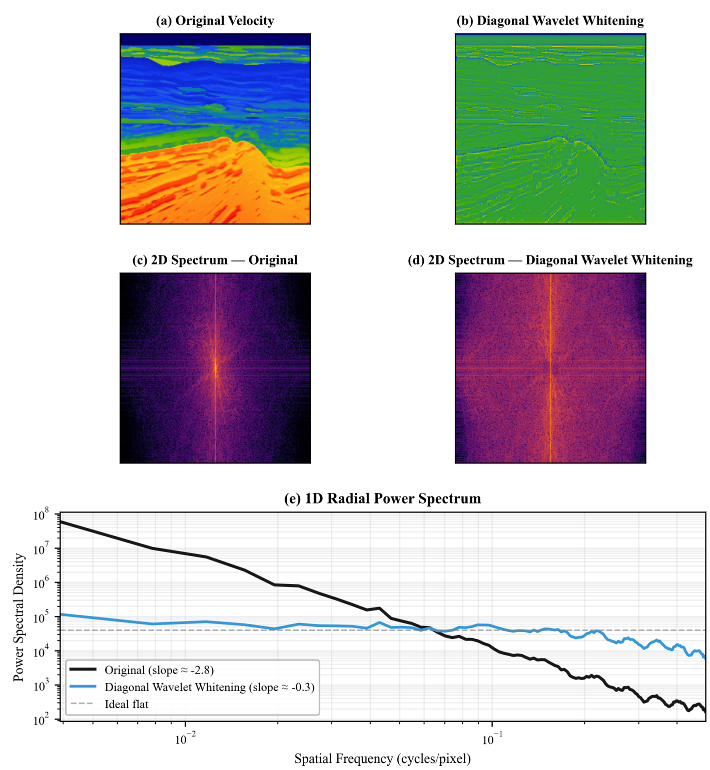

Figure 2: Whitening effect. (a) Original velocity; (b) after whitening. (c, d) 2D Fourier spectra. (e) Radial power spectrum: slope flattens from -2.8 to -0.3.

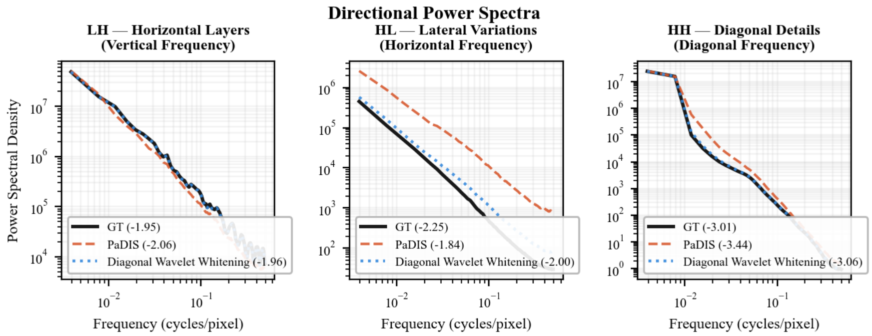

Figure 3: Directional power spectra. LH (horizontal layers), HL (lateral variations), HH (diagonal details). PaDIS (orange dashed) deviates from GT (black solid); Diagonal Wavelet Whitening (blue dotted) reduces the mismatch.

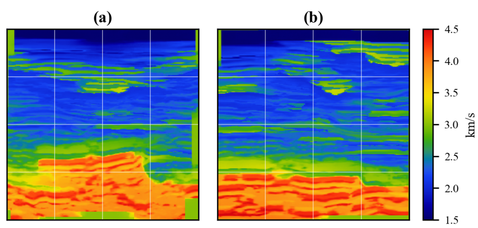

Figure 4: Generated velocity models. (a) PaDIS; (b) with Diagonal Wavelet Whitening and boundary-aware sampling. White grid marks 64-pixel patch boundaries.

Boundary-aware sampling and seam reduction

The hard write-back in standard PaDIS imprints a fixed seam lattice whose locations repeat at every diffusion step, and at low step counts the stationary seams cut through geological layers (Figure 5a, detail crop). Our boundary-aware sampler replaces fixed shifts with step-dependent Halton offsets and hard write-back with cosine-windowed fusion (Section 2.3). We quantify seam severity with the metric $r_{\text{seam}} = \text{MSE}_{\text{boundary}}/\text{MSE}_{\text{interior}}$, the ratio of mean squared pixel difference at patch boundaries to that at non-boundary locations; $r_{\text{seam}} = 1$ indicates no visible seams, while values above 1 indicate boundary discontinuities. At equal compute (48 denoiser evaluations per sample), r_{seam} drops from 1.22 (PaDIS) to 1.05 (proposed), a 14% reduction (Figure 5). Suppressing seams before dewhitening is essential: the inverse whitening amplifies any residual boundary artifact back into velocity space, so eliminating seams at their source preserves geological continuity in the recovered samples.

Computational efficiency and 3D scalability

Patch-based training decouples model capacity from image resolution: with P=64 patches on 256×256 images, the network processes 64×64 crops—16× fewer pixels per forward pass—and GPU memory scales with P^2. The proposed whitening adds negligible overhead, computed once as a preprocessing step; the dominant cost remains the patch-level UNet evaluations, identical to standard PaDIS. A full-image WSGM model on 256×256 data requires approximately 7.9 GB GPU memory (Cirakman et al., 2025); our patch-based model trains within

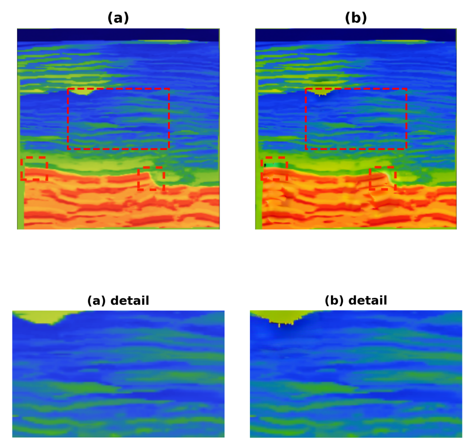

Figure 5: Sampler comparison (48 denoiser evaluations). (a) Standard PaDIS ($r_{\text{seam}} = 1.22$); (b) boundary-aware sampling ($r_{\text{seam}} = 1.05$, 14% reduction). Detail crops show discontinuities in (a) absent in (b).

approximately 4 GB. For three-dimensional volumes of size 256^3, full-image training is infeasible on current hardware, but 64^3 patch training remains within single-GPU budgets.

CONCLUSIONS AND OUTLOOK

We introduced Diagonal Wavelet Whitening and boundary-aware sampling as modular additions to PaDIS for patch-based velocity generation. Whitening reduces the radial spectral slope from -2.8 to -0.3 with exact invertibility, accelerates training by 50%, and improves directional spectral fidelity across all orientations (LH: 91%, HH: 88%, HL: 39% mismatch reduction relative to PaDIS). Boundary-aware sampling reduces the seam metric r_{seam} by 14% without post-processing. Because both the wavelet transform and diagonal standardization are separable, the full pipeline extends to three-dimensional seismic volumes with linear cost scaling and no architectural changes—making it a practical route toward 3D posterior sampling in full-waveform inversion (Cirakman et al., 2025).